A Sequence-to-Sequence Model Approach for ImageCLEF 2018 Medical Domain Visual Question Answering

Rahul Ambati, Student Member IEEE
Multimodal Perception Lab
International Institute of Information Technology, Bangalore
Bangalore, India
Rahul.Ambati@iiitb.org

Chakravardhan Reddy Dudyala
Multimodal Perception Lab
International Institute of Information Technology, Bangalore
Bangalore, India
Chakravardhan.Reddy@iiitb.org

***Abstract*— Numerous attempts have been made in the recent past for the task of free-form and open-ended Visual Question Answering (VQA). Solving VQA problem typically requires techniques from both computer vision for a deeper understanding of the images and Natural language processing for understanding the semantics of the question and generating appropriate answers. It has caught the attention of a lot of researchers because of its enormous applications in the real-world scenarios. But none of the existing approaches are designed for the medical image-question pairs which require a sequence of words as an answer. We propose a novel approach by combining the tasks of Image captioning and Machine translation and provided a comprehensive model that takes a medical image-question pair as an input and generates a sequence of words as an answer. We evaluate our model on the dataset provided by ImageCLEF as a part of the ImageCLEF 2018 VQA-med challenge. We outperformed all the contestants of the challenge by achieving the best BLEU and WBSS scores. Furthermore, we provide additional insights that can be adopted to develop our baseline model and the challenges that lie ahead of us while building Machine learning models for medical datasets.**

Keywords—Visual question answering, ImageCLEF VQA-med, Image captioning, Machine translation, Computer vision, Natural Language Processing

I. INTRODUCTION

Previously the task of processing and classifying an image automatically was very challenging for the humans. But with the recent advancements in Deep learning and Image processing, these tasks have become within reach [1]. There has been a significant amount of research going on in the field of Artificial Intelligence (AI) and deep learning. Some of the inter-disciplinary tasks like Image captioning are the ideal examples for showing the advancement of the multi-discipline deep learning models. Also, Image techniques like segmentation and object recognition have shown us that active research is going on in deeply understanding the image. Our goal is to take forward these existing state-of-the art models for Image captioning and Machine translation and to come up with a comprehensive model for End-to-End task of Visual Question Answering (VQA) [1]-[6].

What makes Visual Question Answering an End-to-End task? In Image captioning the model understands the high-level features of the image and provides captions. But these captions might not give the actual understanding that the humans desire from the image. Whereas in the VQA system takes input an image and a free-form, open-ended, natural-language question about the image and produces a natural language as the output. In the VQA system, the users get what they desire to understand from the image. Hence VQA has emerged as a prominent inter-disciplinary research problem both in academia and industry. It is a fairly complex task as it lies in the intersection of computer vision, natural language processing and deep learning [8]. To accurately answer the question related to the image, the model needs to deeply understand both question and the image. This is a challenging task because historically these both fields have used distinct methods and models to solve their respective tasks [7]. Even though with a very short history, it already has received great research attention from machine learning community. This task has various applications especially for the visually impaired people to help them understand their surroundings. In the medical field, it helps doctors in accurately identifying the tumors or infections from the CT, MRI and PET scanned images.

Recently with the advancements in Machine learning models and consistent improvement in accuracies, Artificial intelligence has paved its way into the medical domain especially in analyzing the medical images. People have started using AI to support the clinical decision making and enhance patient engagement. VQA models have provided a way to significantly enhance the Doctors confidence in interpreting complex medical images by providing a 'second opinion'. For example, consider the following image-question pair: an image showing the CT scan of human lungs and a question "What does the CT scan demonstrate?". Answering these kinds of questions would require VQA model to first understand the semantics of the question, then to locate the objects (lungs) in the image, understanding the relations between the image objects (lungs and infections) and to finally identify the infection and generate an answer. If an automated system does this kind of work, it helps doctors to confirm the infection without any second thoughts.

So far, all the existing approaches and models for VQA system have concentrated on providing a single word answer (mostly 'yes' or 'no) [7]. Though in the recent times sequence to sequence modelling has grown pace, till now there is no comprehensive model which provides sequence of word answers for the VQA task. This could be because of the lack of the dataset that serves as a benchmark for question-

978-1-5386-8235-7/18/$31.00 ©2018 IEEE

answering on the real-world images with multiple word answers. Also, since most of the questions about images often tend to seek specific information, simple one-to-three-word answers are sufficient for many questions [2]. On the other hand, medical questions require a greater number of words to generate an understandable answer which makes the VQA task fairly complex. Also, training the Machine learning models on the medical images often results in lower accuracies because of the complexity involved in those images. Hence there is a need to develop a comprehensive model which takes in an input medical image-question pair and generates a sequence of words (~7) as an answer.

In this paper, we propose a novel approach for the task of Visual Question Answering on the Medical dataset. Unlike the existing models, our model will generate a sequence of words as an answer to the medical image-question pair. Our model took birth from a simplistic idea of combining both Image captioning and Machine translation. Image Captioning involves deeper understanding of the image to provide them captions and Machine translation involves deeper understanding of the input sequence to provide a translated sequence. We tweaked these two models and united them to provide a comprehensive VQA model which generates sequence of words as answers. Our model is trained on the dataset called ImageCLEFmed provided by the ImageCLEF as a part of the challenge named 'ImageCLEF 2018 VQA-Med' organized by CrowdAI. ImageCLEF is the image retrieval track of the Cross-Language Evaluation Forum (CLEF). ImageCLEFmed is a part of ImageCLEF focusing on medical images [12]. Over seventy groups registered for the VQA-med challenge and Obtained access to the datasets. Our model outperformed all the submissions in both BLEU and WBSS metrics. This paper is organized as follows. In section 2 we discuss the existing VQA methods and models. In section 3 we describe the ImageCLEFmed dataset in more detail. In section 4 we describe our model and in section 5 we evaluate our model, report the accuracies and provide future works. We conclude our paper in section 6.

II. Related Works

Our paper is fairly inspired by the success of Image captioning and Machine translation models in the recent times. VQA has gained great research attention recently and a lot of researchers have attempted to solve this problem. The first feasible solution to VQA problems was provided by Malinowski and Fritz [10]. They have used sematic language parser and a Bayesian reasoning model, to understand the semantics of questions and to generate answers. Malinowski and Fritz are the first person to construct a VQA benchmark dataset, named as DAQUAR, which contains 1449 images and 12468 questions generated by humans or automatically. Later, Ren et al. [6] released the TORONTO-QA dataset, which contains a large number of images (123,287) and questions (117684), but the questions are automatically generated and thus can be answered without complex reasoning. Recently, Antol et al. [2], published the currently largest VQA dataset which consists of images from Microsoft COCO dataset. All these datasets have helped the researchers to perform rigorous evaluations on their models.

Most of the existing VQA models employ deep neural models that predominantly uses Convolutional neural networks (CNNs) to extract image features and Long-short term Memory network to extract the semantic representation of the questions because most of these approaches have already gained wide popularity due to their state-of-the-art performance. There are some non-deep learning approaches which have been attempted by the researches for the VQA task. Answer Type Prediction (ATP) [14], is one such kind where the authors propose a Bayesian framework in which they predict the answer type for a question and use this to generate answers. Zhou et al. [15] in 2015 proposed a baseline model called iBOWIMG for VQA. They used pre-trained GoogleNet model and concatenated the image and text features and performed SoftMax regression across the answer classes. Ma et al. [16] propose a CNN only model where they use two CNNs for images, questions and a join CNN to combine the image and question encoding together. Recently in 2016 Malinowski et al. (Ask your Neurons) [1] used CNNs for image and LSTM or GRU for the questions. The answer can be decoded either by classification or by a decoder LSTM. Ren et al. [6] proposed a very similar model to 'Ask your Neurons' model. They used VGGnet for the image and LSTM for the question encoding. In contrast to the previous model, they provided encoded image as the first word to this LSTM network before the question. Noh et al. [17] introduced new technique called Dynamic Parameter prediction for the VQA task. They took VGG architecture, removed final SoftMax layer and added three fully connected layers followed by SoftMax over answer classes. The Second FC layer doesn't have fixed parameters instead they use parameters from the GRU network which is used to encode questions. Recently people have started using attention-based models for both Image captioning and Machine translation to focus on the important parts of the image and text respectively. Shih et al. [18] proposed an attention-based model for VQA called 'Where to Look'. After finding image encoding an attention vector is computed over the set of image features to decide which region in the image is important for the given question. The final image representation is the attention weighted sum of the different regions. Yang et al. [19] have proposed similar model named 'stacked attention networks'. They used the attention weighted image and concatenate with the question encoding. This is used to again compute attention over the original image. They said that this type of 'stacked' attentions helps model to iteratively discard unimportant regions in the image. Lu et al. [11] have proposed a Co-attention model where in addition to modelling visual attention, it also models question attention to focus on the important part of the question. Most of the above-mentioned models uses BLEU evaluation metric. Since these models don't solve the problem of question answering in medical domain, we can't compare our model's performance with them.

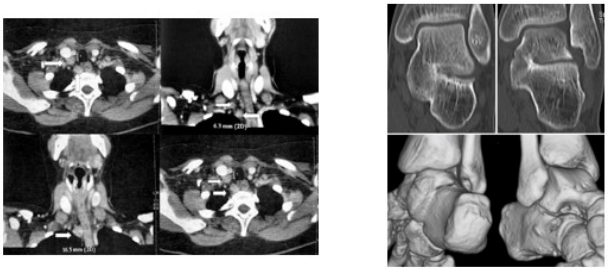

Fig 1. Examples of compound figures in ImageCLEFmed dataset.

III. DATASET

The dataset we used for our model is ImageCLEFmed dataset provided by ImageCLEF as a part of the challenge. The dataset contains 2278 images for training set and 324 images for validation set. The training data consists of 5413 question-answer pairs on these 2278 images. The validation data consists of 500 question-answer pairs on the 324 images. The test set contains roughly 257 images and 500 questions for which our model has to predict the answers. All the questions are content based querying specific information from the image. Most of the images are black and white and are of roughly 700 x 700 (approx.) resolution. They are either CT scans or MRI scans of different parts in the body. Some of the images are compound or multi panel images [12]. Making the content of the compound figures accessible for the model can improve accuracies. Hence the detection of compound figures and their separation into sub figures is an important and complex task for the model. Example of the compound figures are shown in the fig 1.

Most of the questions are either asking whether any infection or tumor is present or seeking more information about the infection. So, the average length of the answer is approximately 5.5 words. Figure 2 shows the frequency of various length answers.

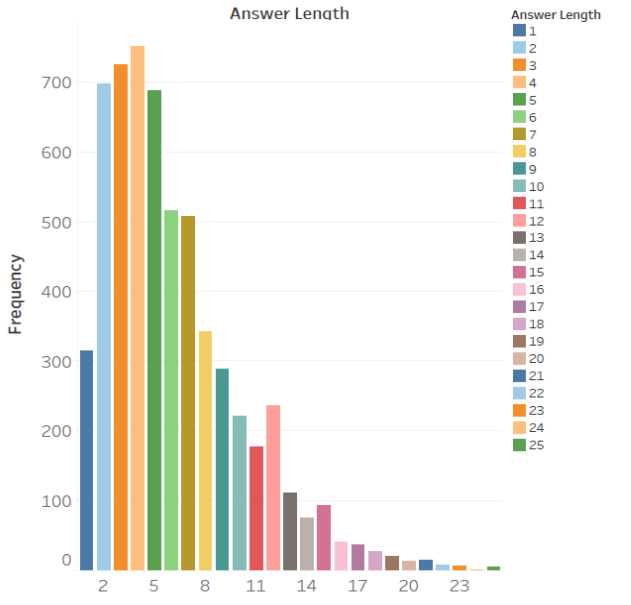

Fig 2. Histogram showing the frequency of various length answers.

IV. METHODOLOGY

The steps adopted in the methodology along with flowcharts depicting the model are listed in this section.

A. Overview:

For the task of VQA we have developed a neural network that takes question embeddings and combines it with the corresponding image embeddings to predict a sequence of words that form a sentence which is the answer to that particular question. Visual Question answering for a single word answer can be formulated as a parametric probability measure:

$$\hat{a} = \arg max_{a \in V}\ p(a|x, q; \theta) \quad (1)$$

where $\boldsymbol{a}$ is the answer prediction by the model, $\boldsymbol{x}$ is the input image, $\boldsymbol{q}$ is the question, $\boldsymbol{\theta}$ is all the parameters of the model that are learned from the training data and $\boldsymbol{V}$ is the complete vocabulary set containing all the words in the questions and answers from training data. The word with the maximum probability predicted by the model is $\hat{\boldsymbol{a}}$ which is most likely to be the answer to the question. Here, the question $\boldsymbol{q}$ is a sequence of words where each word belongs to the vocabulary $\boldsymbol{V}$.

In the case of multiple word answers, the model needs to predict a sequence of words $a = \{a_1, a_2, \dots, a_n\}$ where each word $a_{t\in(1,n)}$ belongs to the vocabulary $\boldsymbol{V}$. The problem that arises here is that the sequence of predictions $a_1, a_2, \dots, a_n$ should form a meaningful sentence and the length of each answer could be varying. Unlike all other existing methods as discussed in previous sections, this task of VQA has an additional constraint of predicting answers that could be as short as 1 word or can even be as long as 20 words.

To tackle this problem, we have added two extra tokens *#start#* and #end# to the set $\boldsymbol{V}$. For each of the answer sequence in the training data, the start token- *#start#* is added in the beginning and the end token *#end#* is added in the end. By adding these tokens, the model learns that the answer sequence always starts with the start token and ends with the end token. During prediction, which happens sequentially, i.e. the model continuously predicts each word $\hat{a}_i$ by taking the previous predictions $\hat{a}_1, \hat{a}_2, \dots, \hat{a}_{i-1}$ into consideration and terminates the prediction process for that question when it encounters an end token; implying that there is a high probability of ending the sequence at that word prediction when $\hat{a}_i$ = *#end#*. Since the predicted sequence also has the end token, we ignore the last word of the prediction and consider the answer as $\hat{a} = \{\hat{a}_1, \hat{a}_2, \dots, \hat{a}_{i-1}\}$. Thus, the probability measure for the task of multiple word answer VQA would be:

$$\hat{a}_i = \arg max_{a \in V}\ p(a|x, q, \{\hat{a}_1, \hat{a}_2, \dots, \hat{a}_{i-1}\}; \theta) \quad (2)$$

B. Method

We have built a deep neural network using (Convolutional Neural Networks) CNNs for obtaining image embeddings and (Gated Recurrent Units) GRUs for both the encoder and decoder models. Since question and answer words cannot be directly fed into the network, they have been first represented as a matrix and later represented as embeddings in higher dimension. These word embeddings are passed to the encoder model. Image embeddings are obtained from CNN model. Once the output of question encoder and image embeddings are obtained, it is combinedly used as the initial hidden state for the decoder model which takes in word embeddings at each time stamp of the answer sequence as input and predicts the word in the next time stamp. While the model is run for several epochs all the parameters in CNN, GRU layers, the latent hidden representation, and the word embeddings are learnt. While testing the decoder predictions at particular time stamp are fed again into the model as input in the next time stamp. A detailed architecture of the model is shown in Fig 3 which is similar in some aspects to Malinowski et. al., (2016) [1].

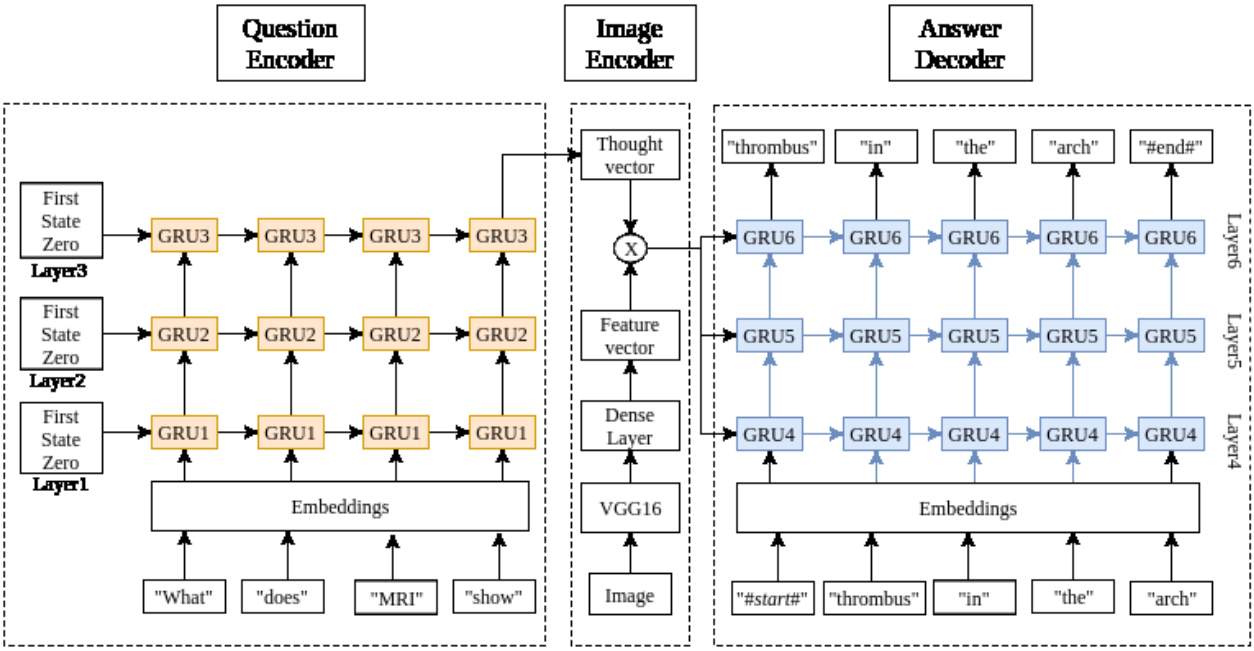

Fig 3. Architecture of the VQA model.

C. *Question encoder*

The vocabulary set V is formed by considering all the unique words in the whole corpus (questions and answers) in the training data. Each of the word is given a unique index value which is a direct word to number mapping. But this representation does not have the information about the context or the semantic relation between the words. To overcome this problem the index representations are passed on to the Keras embedding layer to create embeddings of size 128. Although these embeddings are not precise in the first pass, they continuously get updated with each pass while training. Since the input is a sequence of embeddings and is recurrent in nature, we investigated a few recurrent models like the Long-Short Term Memory(LSTM) [18], Gated Recurrent Unit(GRU) [19] and found out that the GRU network is giving better results and is faster when compared to other recurrent models. GRUs have a simpler architecture than the LSTM cell as it does not have a memory unit [19] and exposes the complete hidden content without any control, which could be very helpful if the model need not predict very long answers. Since most of the answers and questions are short in length, we have chosen GRU for both the encoder and decoder model. The encoder model has 3 GRU layers each with an output dimension of 1024. Each of the word embeddings are sent to the encoder model at each time stamp. The final output after all the embeddings are passed to the encoder model is a thought vector of dimension 1024 which capture the meaning of the question. For the task of sequence to sequence prediction like machine translation etc., this thought vector is used as the initial state for the decoder model. But, in this case we also need to include the image features which is discussed in later section.

D. *Visual Encoder*

The second important component of the model is the visual encoder which converts all the images of various dimensions to a single feature vector. Every image in the dataset is resized to 224x224 pixels. Since the dataset is comparatively very small-2278 images , building a deep convolutional neural network to extract the features would not be viable. Instead, we can use pre-trained network like the VGG16 [20] which a 16-layer network is used by the VGG team in the ILSVRC-2014 competition. The last layer of the network which is essentially the SoftMax layer for prediction is removed. All the other layers in the network are frozen. Each of the pre-processed image in the ImageCLEF VQA dataset is passed through the fine-tuned VGG16 model which gives out a 4096-dimensional feature vector. These vectors are later combined with the question embeddings to pass on to the decoder model.

E. *Multimodal Embedding*

The question encoder converts the question into a thought vector of 1024 dimension and the image encoder converts the image into a feature vector of 4096 dimension. Multimodal embeddings are obtained by combining feature vector and the thought vector. This combination can be obtained by concatenating, element wise multiplication or summation of both the vectors. Since the dimension of the image vector is very high when compared to the thought vector, an additional dense layer has been applied to the image features to get the dimension of feature vector down to 1024. The weights of this dense layer are learnt while training. Once the image vector has been reduced to 1024 dimensions, it could be combined with the thought vector easily. We performed various experiments to combine both the vectors and found out that element-wise multiplication resulted in faster and better results. Concatenation of both the vectors would become expensive in later stages as this would result in a 2048-dimensional vector which has to be fed in to all three GRU layers of decoder and would increase the number of parameters that the model has to learn. Hence, we have performed element-wise multiplication of the thought vector and the feature vector resulting in a new 1024-dimensional vector. This vector is passed on as the initial state for the three layers of GRU in the answer decoder.

F. *Answer Decoder*

The next component in the model is the answer decoder which predicts the answer for each question. The decoder model has 3 GRU layers. The combined thought vector is given as the initial hidden state to the 3 GRU layers. The method for representing words as embeddings for the decoder is same as that of the question encoder. Each of the words in the answer is represented as an index vector and later passed through an embedding layer which converts it into a 128-dimensional vector. As specified in previous sections a start and an end token are added to each of the answers. Embeddings corresponding to each word is passed on to the model at each time stamp. The model predicts the probabilities of the all the words at each time stamp from which the word with the highest probability is chosen and given as output. While testing, this prediction is passed as input to the model in the next time stamp. This process continues recursively until the model predicts the end token: *#end#*. The model then moves on for the next prediction. The start and the end token are ignored, and the remaining sequence is given as answer by the model.

V. Results and Discussion

In this section, we discuss the results, benchmark, metrics and possible conclusions from the results. The ImageCLEF VQA challenge uses WBSS(Word-based Semantic Similarity) [22] and BLEU [23] score as metrics for its evaluation. Table 1 shows the comparison of our model against other contenders of the challenge(Including post challenge submissions).

TABLE 1. WBSS AND BLEU SCORES OF TOP-5 CONTENDERS

Rank	Participant	WBSS	BLEU
01	Chakri	**0.209**	**0.188**
02	UMMS	0.186	0.158
03	TUA1	0.174	0.135
04	NLM	0.174	0.121
05	Bashar	0.122	0.061

We performed various analysis on the training data to derive conclusions about the model performance. Our findings have been summarized below-

A. *Word Occurrence in training data*

The proposed architecture uses an embedding layer to create the embeddings for each word. Fig 4 shows the number of occurrences of words and their count. For example, there are 1455 words which occurred only twice and 610 words which have occurred only 3 times in the dataset. The graph is skewed towards the y-axis implying that most of the words have very few occurrences in the dataset. This makes it difficult for the neural network to understand and model the less frequently occurring words and hence cannot give appropriate predictions. The ImageCLEF VQA contest allows users to use external dataset. PubMed is one the famous resources for medical journals and books. For better embeddings the given dataset can be merged with the PubMed dataset and techniques like

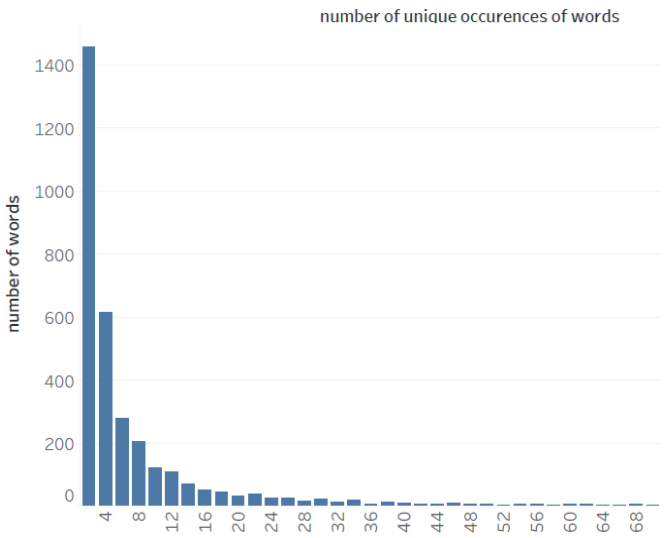

Fig 4. Number of Occurrences of words and the count of such words in the dataset.

the GLOVE [21] can be used. This could give a boost in the performance of the model since word embeddings are the basis for the model to understand the question and predict answers.

B. *Improving Image Embeddings*

The current proposed architecture uses the pre-trained network VGG16 followed by a dense layer for obtaining the image features. We have adopted this architecture because of the small dataset-2278 images. Instead, the ImageCLEF challenge allows participants to use external dataset for training. Additional datasets from the medical domain can be combined with the given dataset to create a bigger dataset which can be helpful to create a deep convolutional neural network. This network would have a better scope in extracting the image features because of the higher number of the training samples. Also, some of the recent advancements in the domain of computer vision are the attention models [8] which can look for some of the specific features in the image.

C. *Variable length Answers*

The distinguishable aspect of this architecture being able to predict variable length answer makes the prediction even more challenging. As seen in the previous section, the answer length varies from as low as 1 word to as long as 25 words. Some of the questions, predicted answers, ground truth are listed below-

- *Question:* What does the abdominal CT scan reveal?

  Answer: A mass in the right lower quadrant.

  Ground Truth: An infiltrating mass.

- *Question*: What structure is seen in the hernia?

  Answer: A large mass.

  Ground Truth: Inflamed appendix.

- *Question:* Is there a mass in the MRI scan?

  Answer: no.

  Ground Truth: no.

- *Question:* What does the MRI show?

  Answer: A round like signal in the upper portion of the right kidney

  Ground Truth: dot-in circle suggesting mycetoma lesions.

It is observant from the predictions that the model can predict very short answers accurately. In the first and the second example, although the prediction seems incorrect, the meaning of the sentence is close to the ground truth. In the last example, the model has predicted a fairly long sentence but couldn't predict the exact sentence. Based on our analysis the model has failed in cases when the answer is very lengthy. Better recurrent units with additional gates to control the length of the sequence could be a breakthrough to this problem.

VI. Conclusion

The field of VQA has matured a lot in the recent times. But still it is taking its baby steps in the field of medicine. In this paper, we propose a comprehensive model for the task of VQA presented by ImageCLEF. The problem with evaluating the existing models on the medical dataset is that most of the models generate answers based on the classification task. But medical questions require sequence of words to generate an understandable answer. Hence the task of VQA needs to be extended to generate sequence of words as an answer. We built a model to modify and combine both Image captioning and Machine translation techniques which are the sub tasks in VQA. We achieved good WBSS and BLEU scores compared to others in the competition 'ImageCLEF 2018 VQA-med'. ImageCLEF conducts participative research and has shown an important impact in visual medical information retrieval. This competition is first of its kind because VQA itself is a new task evolving. Our model is by far the best performing model on the ImageCLEFmed dataset. Novel ways of computing attention improved the performance of both computer vision and Natural Language processing techniques. In future, models from that space can be used as a guide to VQA models. It would be interesting to explore answering as a sequence generation task as more complex questions require greater number of words to answer.

Acknowledgment

We thank ImageCLEF and CrowdAI for providing the ImageCLEF VQA dataset. We thank Prof. Dinesh Babu for guiding through the research work. We are thankful to International Institute of Information Technology, Bangalore for the infrastructural support.

References

[1] M. Malinowski, M. Rohrbach, and M. Fritz, "Ask your neurons: A deep learning approach to visual question answering," arXiv preprint arXiv:1605.02697, 2016.

[2] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh, "Vqa: Visual question answering," in Proc. ICCV, 2015.

[3] Q. Wu, P. Wang, C. Shen, A. Dick, and A. V. D. Hengel, "Ask me anything: Free-form visual question answering based on knowledge from external sources," in Proc. CVPR, 2016.

[4] H. Gao, J. Mao, J. Zhou, Z. Huang, L. Wang, and W. Xu, "Are you talking to a machine? dataset and methods for multilingual image question answering," in Proc. NIPS, 2015.

[5] C. L. Zitnick, A. Agrawal, S. Antol, M. Mitchell, D. Batra, and D. Parikh, "Measuring machine intelligence through visual question answering," AI Magazine, 37(1), 2016.

[6] M. Ren, R. Kiros, and R. Zemel, "Exploring models and data for image question answering," in Proc. NIPS, 2015.

[7] A. K. Gupta, "Survey of Visual Question Answering: Datasets and Techniques," arXiv preprint arXiv:1705.03865v2, 2017.

[8] I. Ilievski, S. Yan, and J. Feng, "Focused Dynamic Attention Model for Visual Question Answering," arXiv preprint arXiv:1604.01485, 2016.

[9] M. Malinowski and M. Fritz, "A Multi-World Approach to Question Answering about Real-World Scenes based on Uncertain Input," in Proc. NIPS, 2014.

[10] J. Lu, J. Yang, D. Batra, and D. Parikh, "Hierarchical Question-Image Co-Attention for Visual Question Answering," arXiv preprint arXiv:1606.00061v5, 2017.

[11] G. S. d. Herrera, Alba & K.Cramer, Jayashree & D. Fushman, Dina & Antani, Sameer & Müller, and Henning, "Overview of the ImageCLEF 2013 medical task," working notes of CLEF, 1179, 2013.

[12] K. Kafle and C. Kanan, "Answertype prediction for visual question answering," in Proc. CVPR. 2016.

[13] B. Zhou, Y. Tian, S. Sukhbaatar, A. Szlam, and R. Fergus, "Simple baseline for visual question answering," arXiv preprint arXiv:1512.02167, 2015.

[14] L. Ma, Z. Lu, and H. Li, "Learning to answer questions from image using convolutional neural network," arXiv preprint arXiv:1506.00333, 2015.

[15] H. Noh, P. H. Seo, and B. Han, "Image question answering using convolutional neural network with dynamic parameter prediction," in Proc. CVPR, 2016.

[16] K. J. Shih, S. Singh, and D. Hoiem, "Where to look: Focus regions for visual question answering," in Proc. CVPR, 2016.

[17] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola, "Stacked attention networks for image question answering," in Proc. CVPR, 2016.

[18] S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural Computation, 1997.

[19] K. Cho, B. V. Merrienboer, C. Gulcehre, F. Bougares, H. Schwenk, D. Bahdanau, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," in Proc. EMNLP, 2014.

[20] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv preprint arXiv:1409.1556, 2014.

[21] J. Pennington, R. Socher, and C. Manning, "Glove: Global Vectors for Word Representation," in Proc. EMNLP, 2014.

[22] S.A. Hasan, Y. Ling, O. Farri, J. Liu, H. Muller, and M. Lungren, "Overview of ImageCLEF 2018 Medical Domain Visual Question Answering Task," in Proc. CEUR Worskshop, 2018.

[23] K. Papineni, S. Roukos, T. Ward and W.J. Zhu, "BLEU: a method for automatic evaluation of machine translation," in Proc. ACL, 2002.