Project Report on

ImageCLEF 2019 Visual Question Answering in the Medical Domain

Submitted by

K A Siva Vardhan	**B160333CS**
P Satyanarayana	**B160340CS**
B Maheswara Rao	**B160349CS**
R Vamsi Krishna	**B160109CS**

Under the Guidance of

Lijiya A.

Department of Computer Science and Engineering
National Institute of Technology Calicut
Calicut, Kerala, India - 673 601

November 28, 2019

ImageCLEF 2019 Visual Question Answering in the Medical Domain

K A Siva Vardhan   P Satyanarayana   B Maheswara Rao   R Vamsi Krishna

1 Abstract

Visual Question Answering is a model in which questions are answered based on the given image. It is highly evolved in case of general domain. Now in this project we will be focussing on medical domain. Automated systems could help clinicians cope with large amounts of images by answering questions about the image contents. We use UNET architecture to extract image features, and use LSTM to encode the questions. Finally, we concatenate the coded questions with the image features to generate the answers.

2 Introduction

Visual Question Answering (VQA) is a recent and exciting problem at the intersection between Computer Vision (CV) and Natural Language Processing (NLP), where the input is an image and a question related to it written in a natural language and the output is the correct answer to the question. The answer can be a simple yes/no, choosing one of several options, a single word, or a complete phrase of sentence.

From a first glance, the VQA problem seem like a very challenging one. The traditional CV techniques used for extracting useful information from images and the NLP techniques typically used for Question Answering (QA) are very far from each other and the interplay between them seem to be complex. Moreover the ability to construct an useful answer based on such multi-modal input adds to the complexity of the problem. Luckily, the recent advances in Deep Learning(DL) have paved the way to building more robust VQA techniques.

The VQA-Med task was introduced for the first time in 2018, inspired by the open-domain VQA challenges that started in 2015 [2]. Given a medical image and a natural language question about the image, participating systems are tasked with answering the question based on the visual image content. Three datasets were provided for training, validation and testing.

For image processing, we use UNET Architecture. Image features are extracted from the last pooling layer of the UNet architecture. For question processing, we use LSTMs without pre-trained embeddings. Question vectors are extracted from the final hidden layer of the LSTMs. These two are concatenated to get the result.

3 Literature Review

3.1 NLM at ImageCLEF 2018 VQA in the Medical Domain

This paper describes the participation of the U.S. National Library of Medicine (NLM) in the Visual Question Answering task (VQAMed) of ImageCLEF 2018.They studied deep learning networks with state of-the-art performance in open-domain VQA.They selected Stacked Attention Network (SAN) and Multimodal Compact Bilinear pooling (MCB) for their official runs.

3.1.1 Stacked Attention Network(SAN)

- The Stacked Attention Network (SAN) was proposed to allow multi-step reasoning for answer prediction. SAN includes three components: (i) the image model based on a CNN to extract high level image representations, (ii) the question model using an LSTM to extract a semantic vector of the question and (iii) the stacked attention model which locates the image regions that are relevant to answer the question.

- For the image model, we used the last pooling layer of VGG-16 pre-trained on imageNet as image features. For the question model, we used the last LSTM layer as question features. The image features and the question vector were used to generate the attention distribution over the regions of the image.

- The first attention layer of the SAN is then computed to capture the correlations between the tokens of the question and the regions in the image. Multimodal pooling is performed to generate a combined question and image vector that is then used as the query for the image in the next layer. We used two attention layers, as it showed better results in open-domain VQA. The last step is answer prediction. For a set of N answer candidates, the answer prediction task is modeled as N-class classification problem and performed using a one-layer neural network. Answers are predicted using Softmax probabilities.

3.1.2 Multimodal Compact Bilinear pooling(MCB)

- Multimodal Compact Bilinear pooling (MCB) is an attention mechanism that implicitly computes the outer product of visual and textual vectors.

- MCB architecture contains: (i) a CNN image model, (ii) an LSTM question model, and (iii) MCB pooling that first predicts the spatial attention and then combines the attention representation with the textual representation to predict the answers

- For the image model, we used ResNet-152 and ResNet-50 pre-trained on imageNet. For the question model, a 2-layer (1024 units in each layer) LSTM model is used. Concatenated output from both layers (2048 units) forms the input to the next pooling layer. MCB pooling is then used to combine both image and textual vectors to produce a multimodal representation. To incorporate attention, MCB pooling is used again to merge the multimodal representation with the textual representation for each spatial grid location. We also fine-tuned ResNet-50 on modality classification.

4 Problem statement

Given a medical image accompanied with a clinically relevant question, our desired model is supposed to be tasked with answering the question based on the visual image content. This may vary from simple problem such as classification of the image to a complex one such as answer generation.

5 Dataset Collection

Since Visual Question Answering is a large area which contains numerous types of images, countably many questions for each image and also many responses for each question, we restrict our project to medical domain. In the scope of the VQA-Med challenge, three datasets were provided:

- The training set contains 12792 question-answer pairs associated with 3200 training images.

- The validation set contains 2000 question-answer pairs associated with 500 validation images.

- The test set contains 500 questions associated with 500 test images.

The data is equally distributed over four categories based on the question types which are:

5.1 Plane Category

Question on planes come in one of the following formats: “in which plane”, “Which plane”, “what plane”, “in what plane”, “what is the plane”, “what imaging plane is”, and “what image plane”. There are 16 different planes. Some of them are axial, sagittal, coronal, AP, lateral, frontal etc.

5.2 Organ Category

Question on organ systems come in one of the following formats: “what organ system is”, “what part of the body is”, “the ct/mri/ultrasound/x-ray scan shows what organ system”, “which organ system is”, “what organ system is”, “what organ is this”, etc. There are ten organ systems.

5.3 Modality Category

Question on organ systems come in one of the following formats: “what modality was used to take this image”, “is this an mri image”, “is this a t1 weighted, t2 weighted, or flair image”, “what type of contrast did this patient have”. There are eight main modality categories: XR, CT, MR, US, MA, GI, AG, and PT.

5.4 Abnormality Category

Question on organ systems come in one of the following formats: “what is the abnormality/wrong/alarming in this image”, “is this image normal” or “is this image abnormal”.

6 Design

Our model consists of four modules: image feature extraction, question semantic encoder, feature fuse with co-attention mechanism and answer prediction.

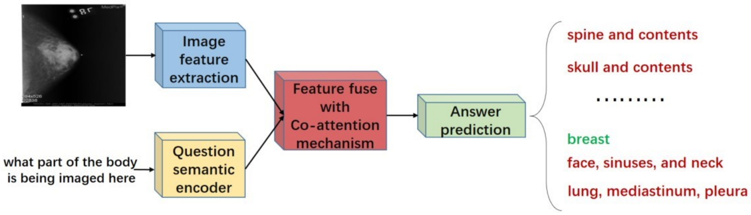


Image feature extraction
what part of the body is being imaged here
Question semantic encoder
Feature fuse with Co-attention mechanism
Answer prediction
spine and contents
skull and contents
....
breast
face, sinuses, and neck
lung, mediastinum, pleura

Figure 1: Our model architecture

6.1 Image feature Extraction

We have used UNET architecture in order to extract image features. There are various levels of granularity in which the computers can gain an understanding of images.

- **Image classification**
  Given an image, we expect to output a corresponding discrete label.
- **Classification with Localization**
  Compute to localize where exactly the object is present in the image.
- **Object Detection**
  Extends classification with localization to next level where the image can contain multiple objects.
- **Semantic Segmentation**
  To label each pixel of an image with a corresponding class of what is being represented.
- **Instance segmentation**
  Expect the computer to classify each instance of a class separately.

The steps involved in UNET architecture are

1. **Convolution operation**
   There are two inputs to a convolution operation
   (a) Input image of size (nin x nin)
   (b) A set of ‘k’ filters (also called as kernels or feature extractors) each one of size (f x f), where f is typically 3 or 5.

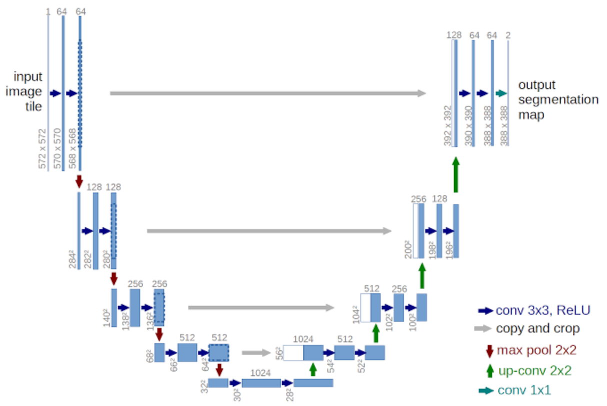


Figure 2: UNET architecture

The kernel strides over the input matrix of numbers(pixels) moving horizontally column by column, sliding/scanning over the first rows in the matrix containing the images pixel values. Then the kernel strides down vertically to subsequent rows. Note, the filter may stride over one or several pixels at a time.

The output of a convolution operation is also an output image or feature map of size (nout x nout x k).

$$n_{out} = \lfloor (n_{in} + 2p - k)/s \rfloor + 1$$

where nin is number of input features, nout is number of output features, k is convolution kernel size, p is convolution padding size and s is convolution stride size(number of movements of kernel function).

2. **Max pooling operation**
   In simple words, the function of pooling is to reduce the size of the feature map so that we have fewer parameters in the network. The idea is to retain only the important features (max valued pixels) from each region and throw away the information which is not important.

   By down sampling, the model better understands "WHAT" is present in the image, but it loses the information of "WHERE" it is present.

3. **Transposed Convolution**
   Transposed convolution (also called as deconvolution or fractionally strided convolution) is a technique to perform up sampling of an image with learnable parameters. Transposed convolution is exactly the opposite process of a normal convolution i.e., the input volume is a low resolution image and the output volume is a high resolution image.

6.2 Question Preprocessing

For a given question, we start with text normalization. Text normalization includes:

- Converting all letters to lower or upper case.
- Removing punctuations.
- Removing white spaces.
- Removing stop words, sparse terms, and particular words.

6.3 Question Semantic Encoder

Long Short Term Memory networks – usually just called "LSTMs" – are a special kind of RNN, capable of learning long-term dependencies. They were introduced by Hochreiter Schmidhuber (1997). LSTMs are explicitly designed to avoid the long-term dependency problem.

LSTMs have this chain like structure, with four neural network layers, interacting in a very special way.

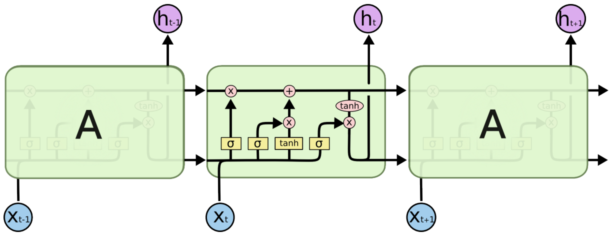


Figure 3: LSTM Networks

The key to LSTMs is the cell state, the horizontal line running through the top of the diagram.The

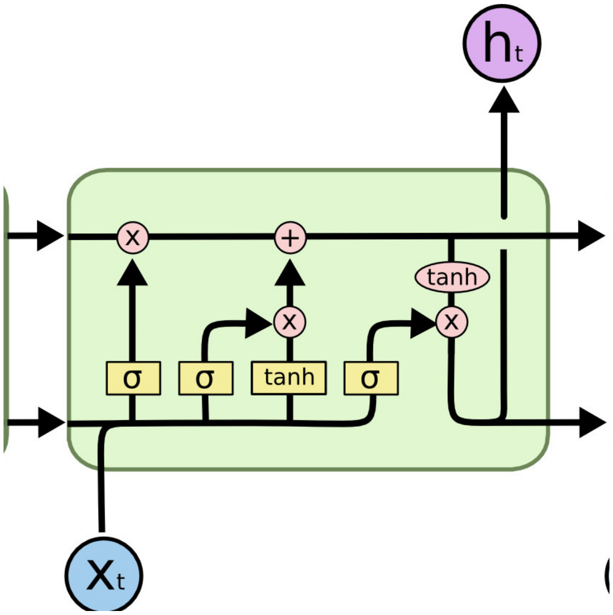


Figure 4: LSTM Networks

LSTM does have the ability to remove or add information to the cell state, carefully regulated by structures called gates.

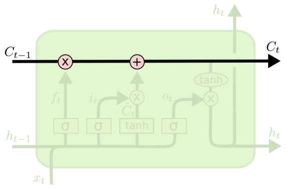
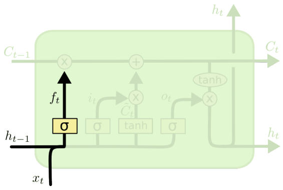


Figure 5: Cell State

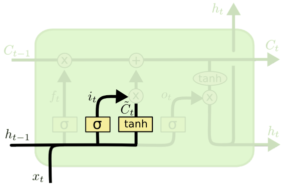
The first step in our LSTM is to decide what information we're going to throw away from the cell state. This decision is made by a sigmoid layer called the “forget gate layer.” It looks at ht-1 and xt, and outputs a number between 0 and 1 for each number in the cell state Ct-1. 1 represents “completely keep this” while 0 represents “completely get rid of this”.

The next step is to decide what new information we're going to store in the cell state. This has two parts. First, a sigmoid layer called the “input gate layer” decides which values we'll update. Next, a tanh layer creates a vector of new candidate values, C‘t, that could be added to the state. In the next step, we'll combine these two to create an update to the state.



$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] \ + \ b_f\right)$$

Figure 6: first layer in LSTM

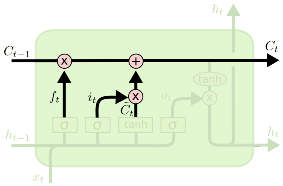


$$i_t = \sigma\left(W_i \cdot [h_{t-1}, x_t] \ + \ b_i\right)$$
$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] \ + \ b_C)$$

Figure 7: second layer in LSTM

It's now time to update the old cell state, Ct-1, into the new cell state Ct. The previous steps already decided what to do, we just need to actually do it.

We multiply the old state by ft, forgetting the things we decided to forget earlier. Then we add it*C‘t. This is the new candidate values, scaled by how much we decided to update each state value.



$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t$$

Figure 8: Calculating Ct in LSTM

Finally, we need to decide what we're going to output. This output will be based on our cell state, but will be a filtered version. First, we run a sigmoid layer which decides what parts of the cell state we're going to output. Then, we put the cell state through tanh (to push the values to be between -1 and 1) and multiply it by the output of the sigmoid gate, so that we only output the parts we decided to.

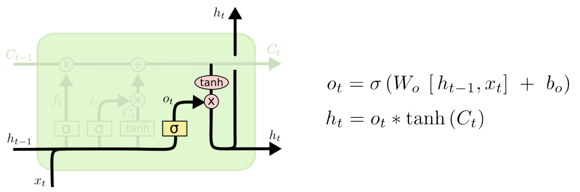

h_t
C_{t-1}
C_t
f_t
i_t
$\tilde{C}_t$
σ
σ
tanh
σ
o_t
tanh
h_{t-1}
h_t
x_t
$o_t = \sigma\left(W_o\ [h_{t-1}, x_t]\ +\ b_o\right)$
$h_t = o_t * \tanh\left(C_t\right)$

Figure 9: Last layer in LSTM

6.4 Answer Generation

Given the image feature matrix vI and the question feature vector vQ. There are two co-attention mechanisms that differ in the order in which image and question attention maps are generated.

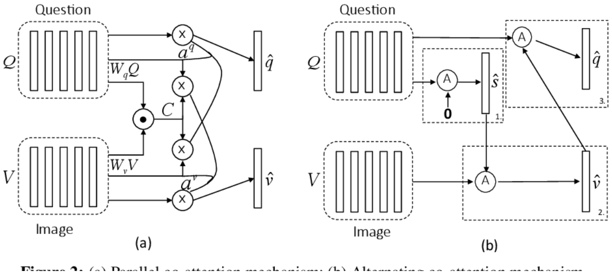


Figure 10: (a) Parallel Co-attention mechanism (b) Alternating Co-attention mechanism

1. **Parallel co-attention**
   Parallel co-attention attends to the image and question simultaneously. Here we connect the image and question by calculating the similarity between image and question features at all pairs of image-locations and question-locations.

2. **Alternating co-attention**
   In this attention mechanism, we sequentially alternate between generating image and question attention. Briefly, this consists of three steps

   (a) Summarize the question into a single vector q

   (b) Attend to the image based on the question summary

   (c) Attend to the question based on the attended image feature.

We predict the answer based on the co-attended image and question features. We use a multi-layer perceptron (MLP) to recursively encode the attention features.

7 Work to be done

We are left with the implementation part which we will be doing in the coming semester. In the implementation we will be processing our data set. As mentioned above we will be using UNET architecture for image feature extraction, LSTM for question understanding and an attention mechanism with Simple multi layer perceptron for answer generation where we concatenate the results obtained from image feature extraction and question understanding. We will seek new models based on the accuracy of the present model and try to increase accuracy as much as possible.

8 Conclusion

We reviewed popular methods in deep learning, and are trying to build a VQA model for ImageCLEF 2019. We will compare our model with the presently accurate model and try to make changes accordingly to increase accuracy. Despite shortcomings of current practices for both training and evaluating VQA systems, we identified a number of promising research avenues that could potentially bring future breakthroughs for both VQA and for the general objective of visual scene understanding. Since deep learning techniques are significantly improving, we can reasonably expect that VQA is going to be more and more accurate in the next years.

References

[1] Y. Zhou, X. Kang, and F. Ren, “Employing inception-resnet-v2 and bi-lstm for medical domain visual question answering,”

[2] A. B. Abacha, S. Gayen, J. J. Lau, S. Rajaraman, and D. Demner-Fushman, “Nlm at imageclef 2018 visual question answering in the medical domain,” 2018.

[3] J. J. Lau, S. Gayen, A. B. Abacha, and D. Demner-Fushman, “A dataset of clinically generated visual questions and answers about radiology images,” Scientific data, vol. 5, p. 180251, 2018

[4] B. Talafha and M. Al-Ayyoub, “Just at vqa-med: A vgg-seq2seq model”

[5] Akira Fukui, Dong Huk Park, Daylen Yang and Anna Rohrbach ”Multimodal Compact Bilinear Pooling for Visual Question Answering and Visual Grounding”

[6] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, Alex Smola. Stacked Attention Networks for Image Question Answering. The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 21-29

[7] Damien Teney, Peter Anderson, Xiaodong He, Anton van den Hengel.Tips and Tricks for Visual Question Answering: Learnings From the 2017 Challenge. The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 4223-4232

[8] Peng Wang, Qi Wu, Chunhua Shen, Anton van den Hengel. The VQA-Machine: Learning How to Use Existing Vision Algorithms to Answer New Questions. The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 1173-1182

[9] Aisha Al-Sadi1, Bashar Talafha1, Mahmoud Al-Ayyoub1, Yaser Jararweh1 and Fumie Costen2 ”JUST at ImageCLEF 2019 Visual Question Answering in the Medical Domain”

[10] Jiasen Lu, Jianwei Yang, Dhruv Batra, Devi Parikh ”Hierarchical Question-Image Co-Attention for Visual Question Answering”