IMAGE CLEF 2019 VISUAL QUESTION ANSWERING IN THE MEDICAL DOMAIN

Under the Guidance of *Lijiya A.*

Group Members:

- K A Siva Vardhan - b160333cs
- B Maheswararao - b160349cs
- P Satyanarayana - b160340cs
- R Vamsi Krishna - b160109cs

TOPICS COVERED

1. Abstract
2. Introduction
3. Literature Review
4. Problem Statement
5. Dataset Collection
6. Design
7. Work to be done
8. Conclusion
9. Reference

ABSTRACT

- Visual Question Answering is a model in which questions are answered based on the given image.
- In our project we will be focusing on medical domain.
- We use UNET architecture to extract image features, and use LSTM to encode the questions.
- Finally, we concatenate the encoded questions with the image features to generate the answers.

INTRODUCTION

This project is about building a model 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.

- For image processing, we use UNET Architecture where Image features are extracted from the last pooling layer of the UNET architecture.
- For question processing, we use LSTMs. Question vectors are extracted from the final hidden layer of the LSTMs.
- These two are then concatenated to get the result.

LITERATURE REVIEW

NLM at ImageCLEF 2018 VQA in the Medical Domain

- Studied deep learning networks with state of-the-art performance in open-domain VQA.
- Selected Stacked Attention Network (SAN) and Multimodal Compact Bilinear pooling (MCB) for their official runs.

Stacked Attention Network(SAN)

- Proposed to allow multi-step reasoning for answer prediction.
- Includes three components
  - Image model based on a CNN to extract high level image representations.
  - Question model using an LSTM to extract a semantic vector of the question.
  - Stacked attention model which locates the image regions that are relevant to answer the question.

Multimodal Compact Bilinear pooling(MCB)

- Is an attention mechanism that implicitly computes the outer product of visual and textual vectors.
- MCB architecture contains
  - CNN image model.
  - LSTM question model.
  - MCB pooling that first predicts the spatial attention and then combines the attention representation with the textual representation to predict the answers.

PROBLEM STATEMENT

- Given a medical image accompanied with a clinically relevant question, our desired model is supposed to answer 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.

DATASET COLLECTION

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.

DESIGN

- Our model consists of four modules:
  - Image feature extraction.
  - Question semantic encoder
  - Feature fuse with co-attention mechanism
  - Answer prediction.

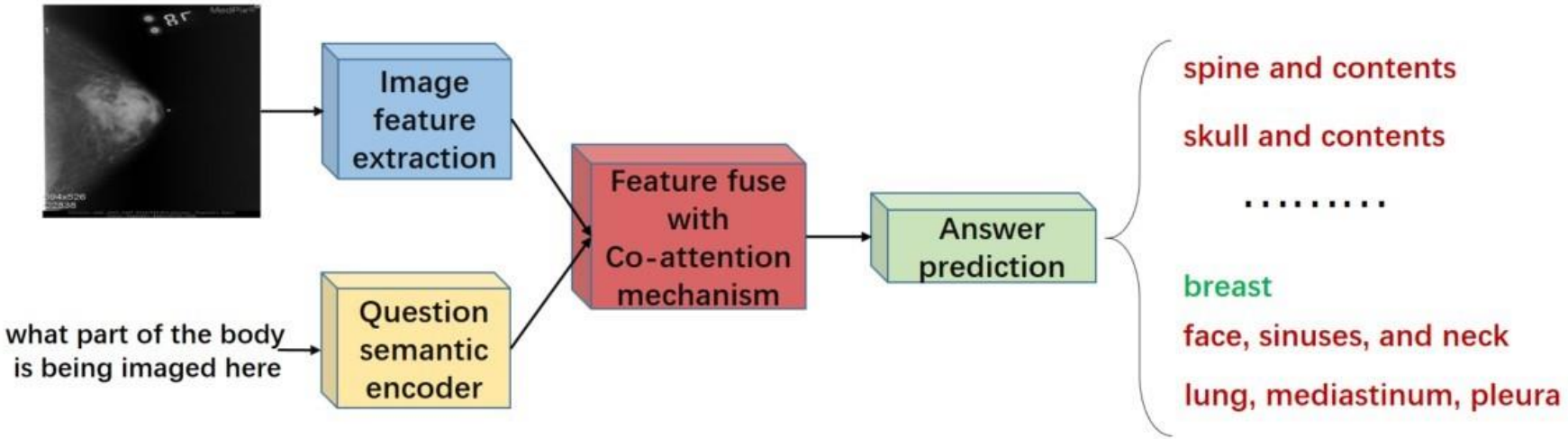


IMAGE FEATURE EXTRACTION

- We have used UNET architecture in order to extract image features.

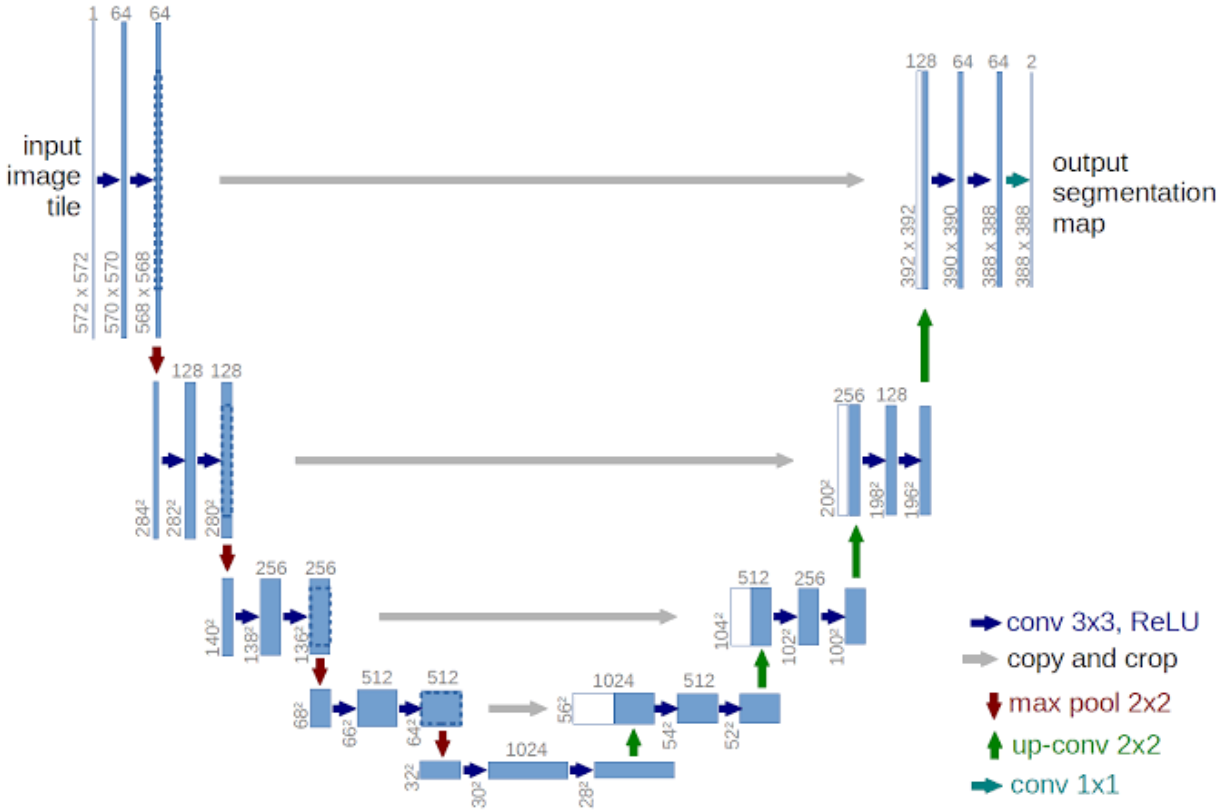


IMAGE FEATURE EXTRACTION

- The steps involved in UNET architecture are
  - Convolution operation
  - Max pooling operation
  - Transposed Convolution

IMAGE FEATURE EXTRACTION

Convolution operation

- There are two inputs to a convolution operation
  - Input image of size (nin x nin)
  - A set of 'k' filters (also called as kernels) each one of size (f x f), where f is typically 3 or 5.
- The kernel strides over the input matrix of numbers(pixels) moving horizontally column by column, sliding 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).
- $nout = (nin + 2p - k)/s + 1$

IMAGE FEATURE EXTRACTION

Max Pooling Operation

- 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 from each region and throw away the information which is not important.

Transposed Convolution

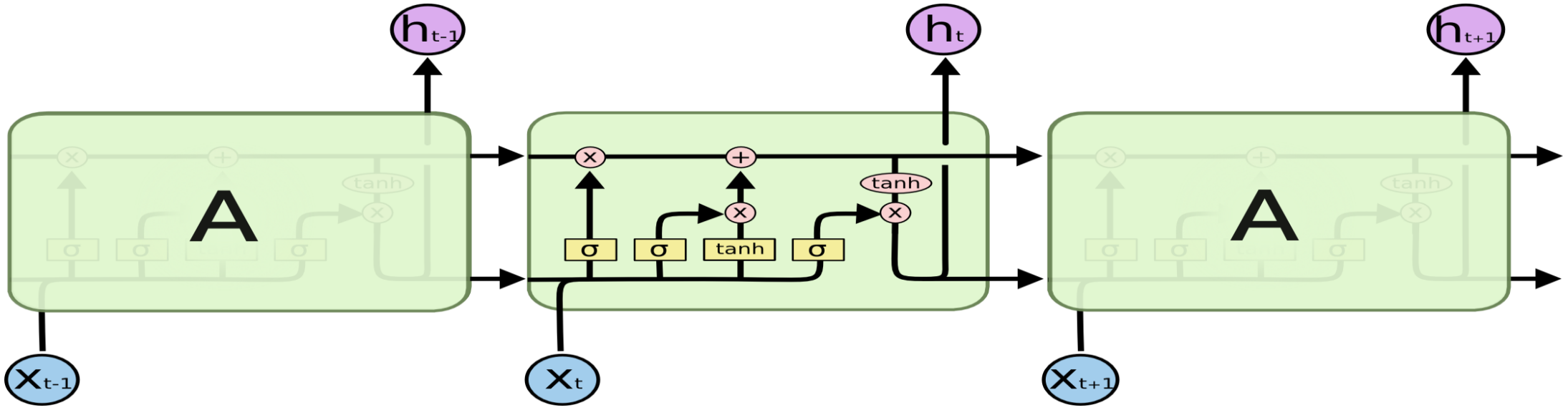
- It is a technique to perform up sampling of an image with learnable parameters.
- It is exactly the opposite process of a normal convolution i.e., the input volume is a low resolution image and the output volume is a high resolution image.

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.

QUESTION SEMANTIC ENCODER

- LSTM(Long Short Term Memory Networks) are a special kind of RNN, capable of learning long-term dependencies.
- LSTMs have this chain like structure, with four neural network layers, interacting in a very special way.

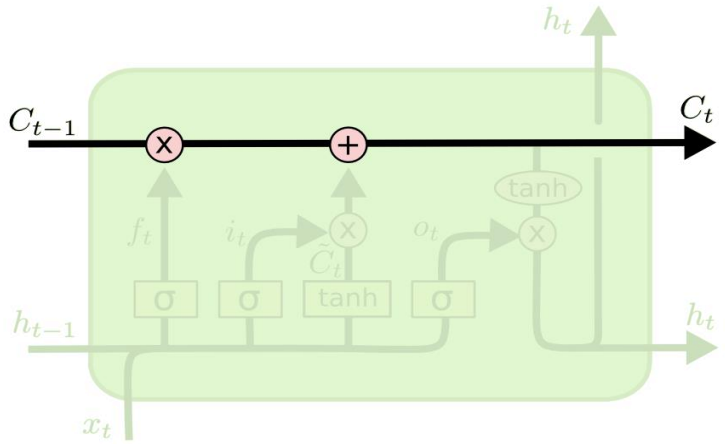


QUESTION SEMANTIC ENCODER

Different layers in LSTM

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

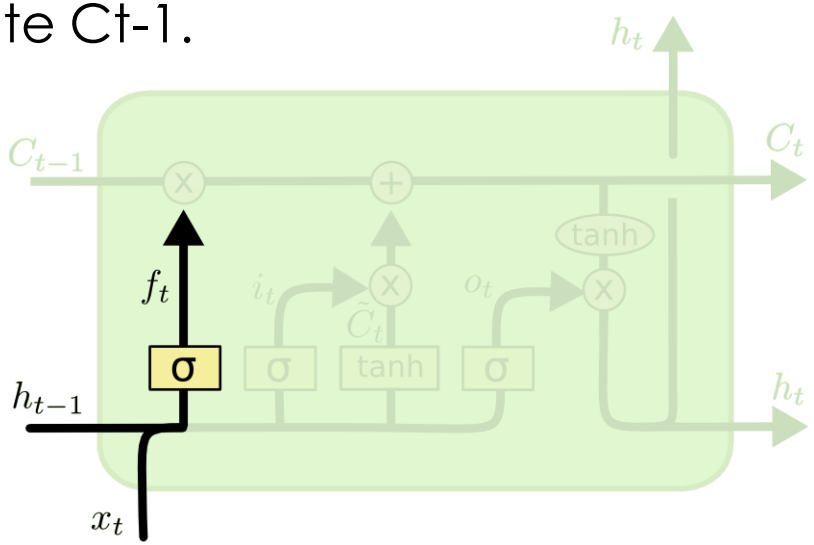
- The key to LSTMs is the cell state, the horizontal line running through the top of the diagram.



QUESTION SEMANTIC ENCODER

Step1

- 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.

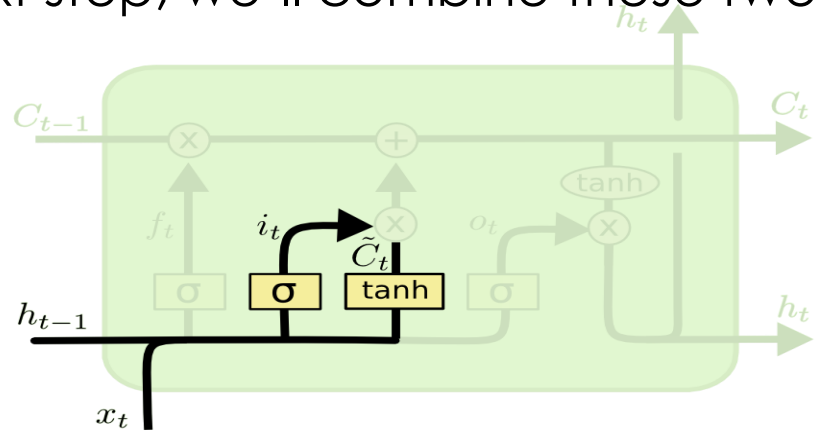


$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] \; + \; b_f\right)$$

QUESTION SEMANTIC ENCODER

Step2

- The next step is to decide what new information we're going to store in the cell state.
- This has two parts.
  - A sigmoid layer called the "input gate layer" decides which values we'll update.
  - 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.



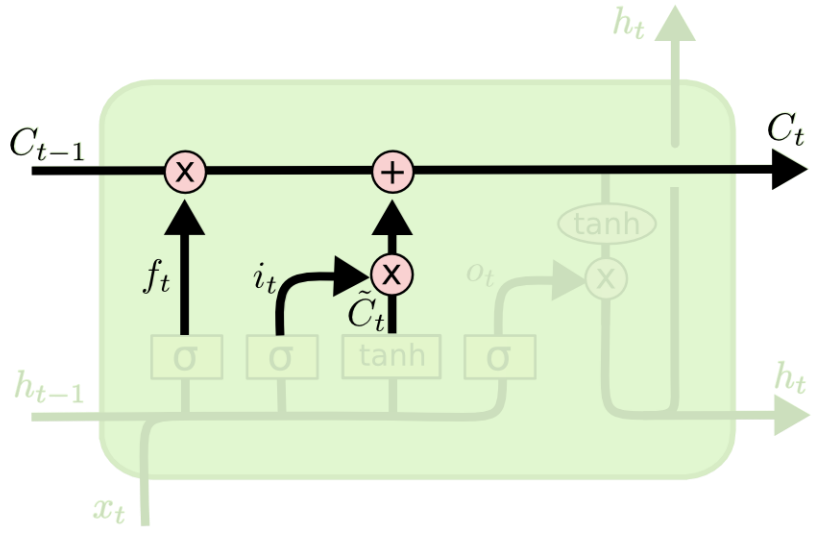
$$i_t = \sigma\left(W_i \cdot [h_{t-1}, x_t] \; + \; b_i\right)$$

$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] \; + \; b_C)$$

QUESTION SEMANTIC ENCODER

Step3

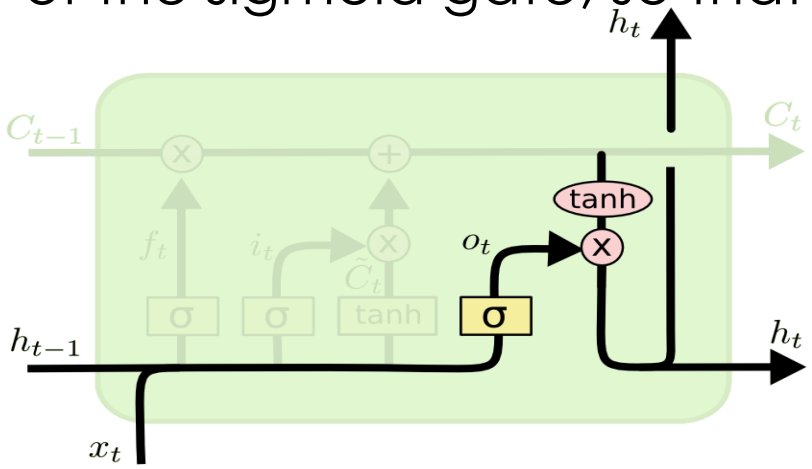
- It's now time to update the old cell state, Ct-1, into the new cell state Ct.
- We multiply the old state by ft, forgetting the things we decided to forget earlier.
- Then we add it to C't. This is the new candidate values, scaled by how much we decided to update each state value.



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

Step4

- Finally, the 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 (range is [-1,1]) and multiply it by the output of the sigmoid gate, so that we only output the parts we decided to.



$$o_t = \sigma\left(W_o\ [h_{t-1}, x_t]\ +\ b_o\right)$$

$$h_t = o_t * \tanh\left(C_t\right)$$

ANSWER GENERATION

- There are two co-attention mechanisms that differ in the order in which image and question attention maps are generated.
  - Parallel co-attention
  - Alternating co-attention

Parallel co-attention

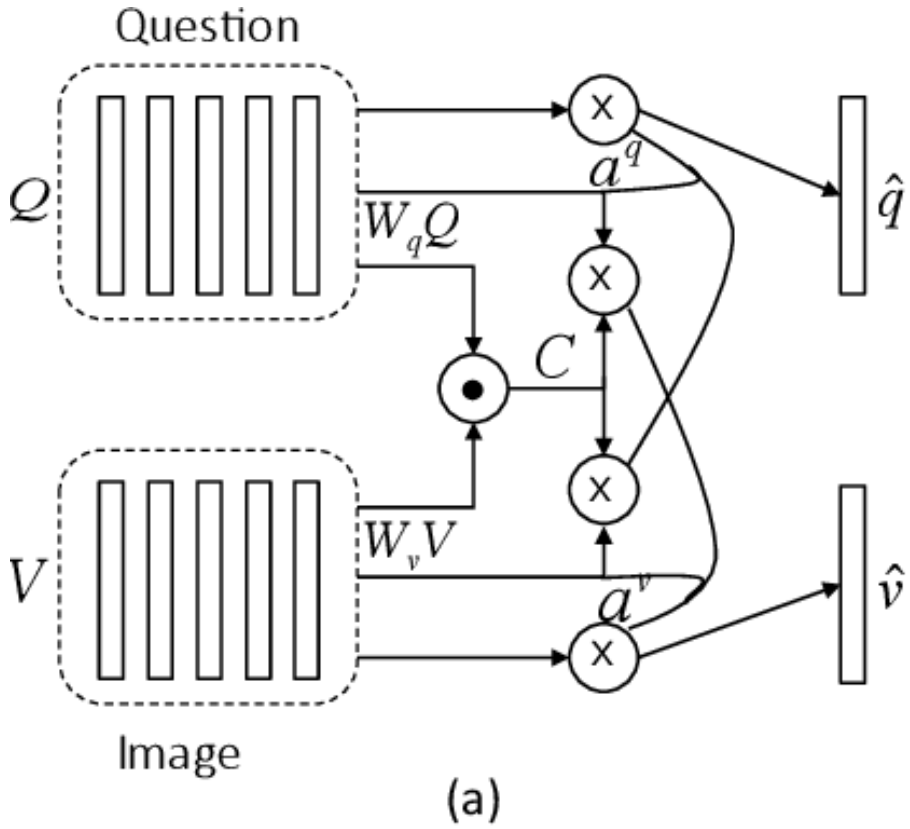
  - Parallel co-attention attends to the image and question simultaneously.
  - Here we connect the image and question by calculating the similarity between image and question features at all pairs of image-locations and question locations.

ANSWER GENERATION

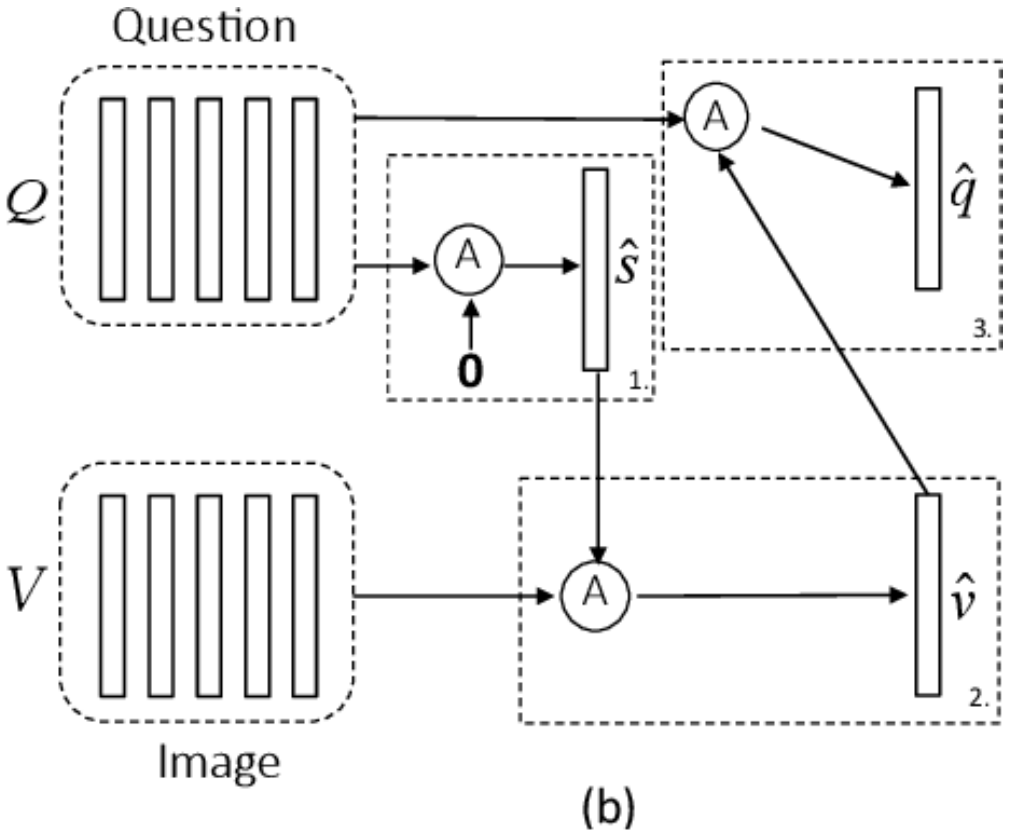
Alternating co-attention

- In this attention mechanism, we sequentially alternate between generating image and question attention. Briefly, this consists of three steps
  - Summarize the question into a single vector q.
  - Attend to the image based on the question summary.
  - Attend to the question based on the attended image feature.

ANSWER GENERATION



Parallel Co-attention mechanism



Alternating Co-attention mechanism

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
  - An attention mechanism with Simple multi layer perceptron for answer generation
- We will seek new models based on the accuracy of the present model and try to increase accuracy as much as possible.

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 in VQA, we identified a number of promising research avenues that could potentially bring future breakthroughs for both VQA and visual scene understanding.
- Since deep learning techniques are significantly improving, we can expect that VQA is going to be more accurate in the next years.

REFERENCE

1. 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 vqamed: 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 AlAyyoub1, 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 CoAttention for Visual Question Answering.”

THANK YOU