Name: MANAS DAS

Email address: manas234das@gmail.com

Contact number: 7008066826

Anydesk address:

Date: **11th Feb 2020**

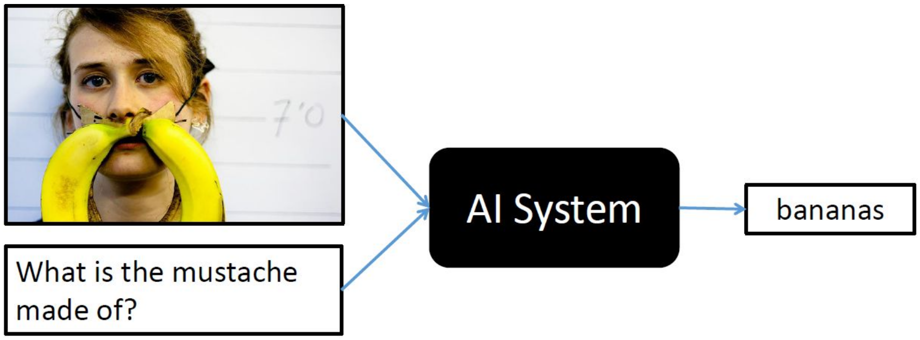
Self Case Study - 1: Visual Question Answering

Overview :

We can define a VQA system as an algorithm that takes as input an image and a natural language question about the image and generates a natural language answer as the output.

1. VQA or visual question answering is a task of free-form and open-ended Visual Question Answering. It is a popular deep learning research which requires coordination of natural language processing and computer vision modules into a single architecture. The challenge is to build an architecture which will not only understand a natural language question but also it will inference the question along with the given image and will try to answer the question as accurately as possible.
2. A system that succeeds at VQA typically needs a more detailed understanding of the image and complex reasoning than a system producing **generic image captions**. The system has to **develop an intuition to understand which specific part of the image might contain the possible answer** and should focus on that specific part.
3. A good VQA system must be capable of solving a broad spectrum of typical NLP and Computer vision tasks, as well as reasoning about image content. It is clearly a multi-discipline AI research problem, involving Computer vision, NLP and Knowledge Representation & Reasoning.

For example :



4. In the above example, we are giving the AI system an image and a natural language question about that image. The task of the AI system is to understand the image and it must produce the correct answer to the question asked.

Research-Papers/Solutions/Architectures/Kernels and First Cut Approach

1. **Problem statement** : Given an image and a natural language question about the image, the task is to provide an accurate natural language answer.

   (For example : Given an image of a child sitting on a table, a natural language question will be asked to the AI system. To that, the model has to answer the question very accurately.)

   Suppose the question asked to the model is

   "Who is sitting on the table?"

   The model should answer : "A child".

   (NOTE : Please refer the image in the previous page)

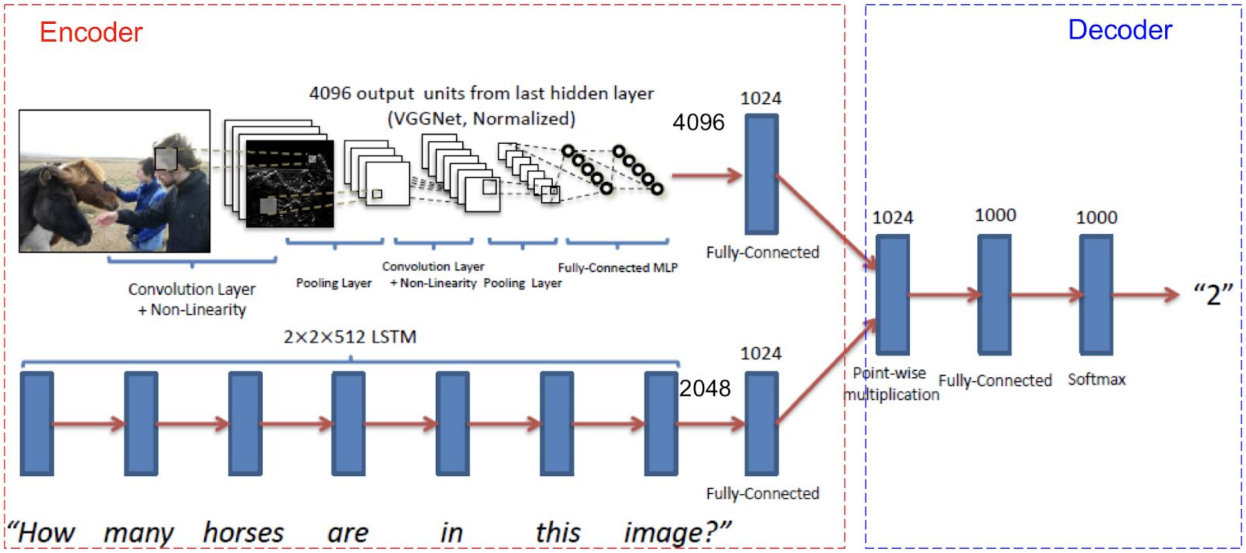
2. **Detailed Description :**

   Methodology:

   There are multiple Models for the task of Visual Question Answering. Two of them are a. Base Model and b. Hierarchical Co-attention Model. A typical system of VQA consists of image, question(represented by text) as inputs and answer to the question as output. Systems differ in how the image and questions features are encoded into a common vector space, followed by decoding the vector space to get the answer. Typically, the image features are computed by Convolution Neural Network(CNN) whereas the text features are computed using Recurrent Neural Network(RNN) to preserve the temporal information in the text. The Base Model considers the aggregate features from question and image to determine the answer. While the Hierarchical Co-attention model determines the answer by attended image and question features.

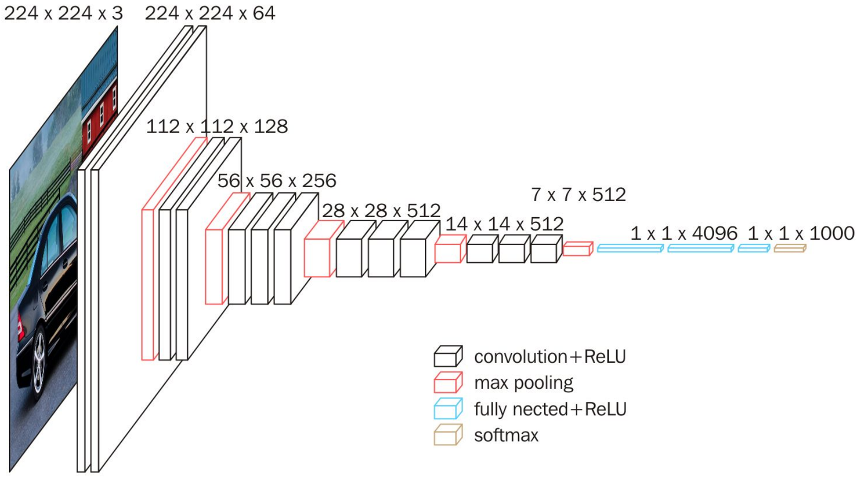
a. BASELINE MODEL : VQA: Visual Question Answering

i. The image and question first needs to be embedded (encoded) into a common vector space and a decoder then decodes the vector space to obtain the answer.



ii. ENCODER : The encoder part consists of image and question encoding.

Image Encoding : A pre-trained vgg16 CNN model on Imagenet is used as an encoder. The Vgg16 model consists of 5 convolution layers, 2 fully connected layers and 1 softmax layer. Output of fully connected layers are considered as image features which are of size 4096.



Question Encoding : RNNs are used to encode the question into vector space by preserving temporal information. LSTM are used as a RNN module to mitigate the problems of vanishing gradient descent. We have a fixed length of LSTM units as we will have a threshold on maximum number of words each question can have. The state of the final LSTM unit is considered as a question feature. A LSTM of 512 unit size is considered in each layer. Each LSTM unit gives hidden state of size 512 and cell state of size 512. Both the states are concatenated to get a 1024 vector. Since two LSTM layers are considered we get a 2048 size vector as a question feature.

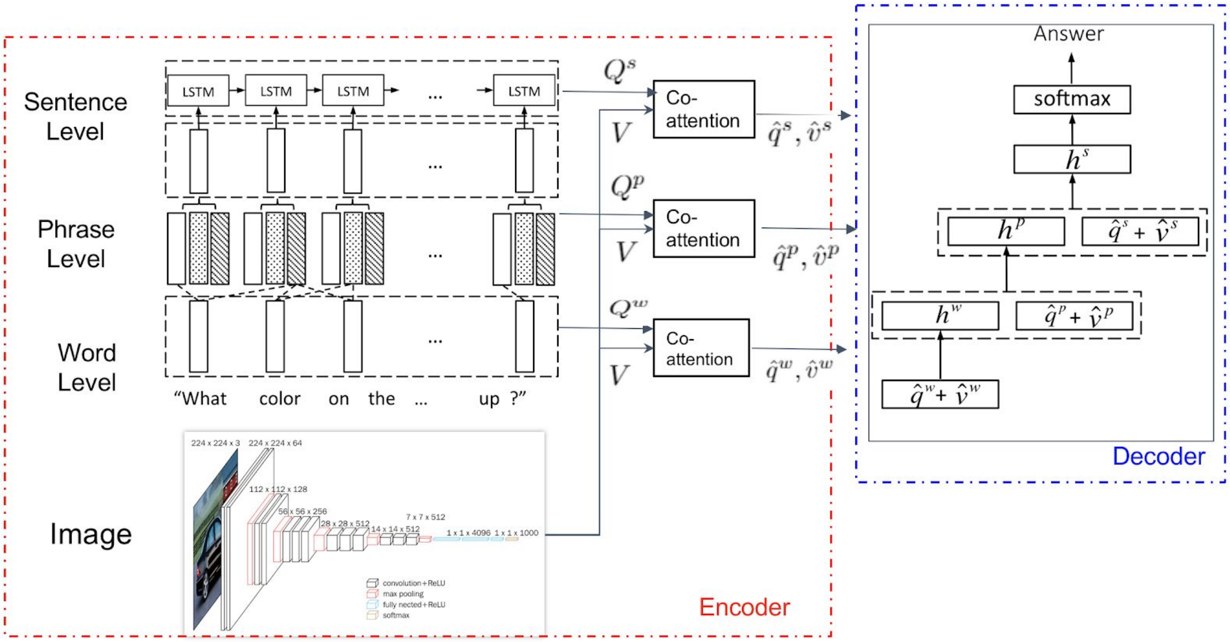
Since the outputs from both image and question encoding are different, we have a fully connected layer at image and question encoding to get them to a size of 1024. Thus the outputs of the encoder are two vectors of size 1024 which represents the image and question features.

iii. DECODER : The Decoder performs a softmax-classification for the image and question features calculated by Encoder. Decoder predicts the best answer among the top 1000 chosen from dataset. The top 1000 answers accommodates around 85% answers of the dataset. Hence this is mostly a classification task rather than generating task for answers. The steps involved in classification are, First, a pointwise image and question features are multiplied to get a single vector of size 1024. This is fed to a fully connected layer of size 1000 and softmax layer. The highest output from the softmax layer is the answer to the given question.

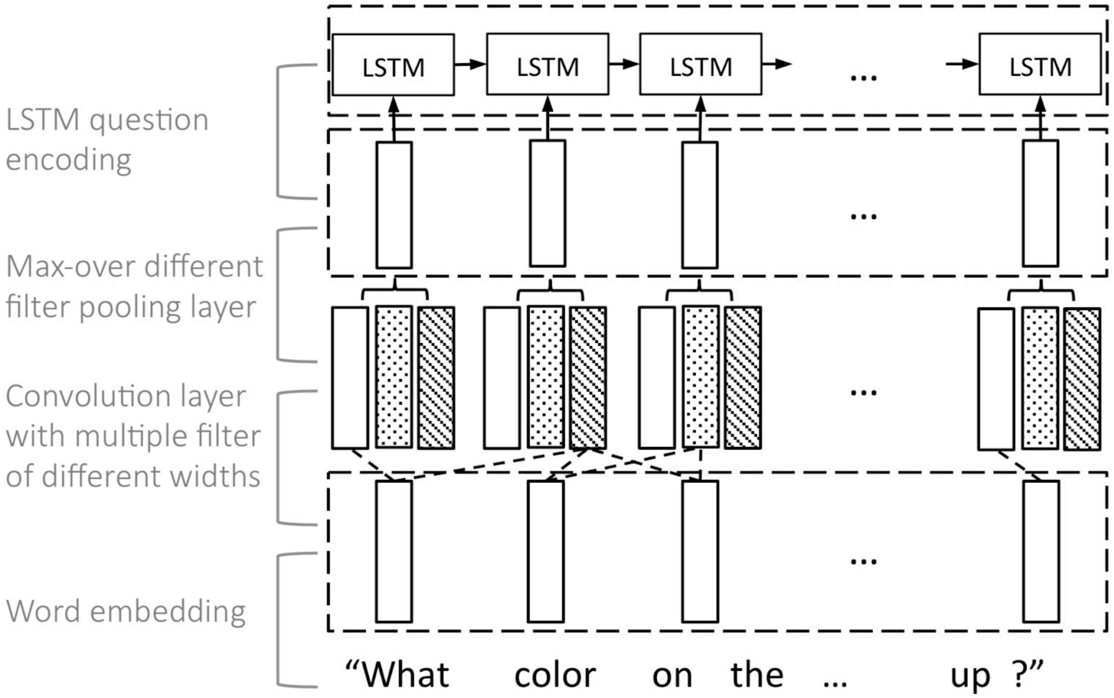
GitHub page : https://github.com/GT-Vision-Lab/VQA_LSTM_CNN

REFERENCE : https://github.com/Heronalps/Visual_QA_Attn

b. Hierarchical Co-Attention Model : Hierarchical Question-Image Co-Attention for Visual Question Answering



i. In the Base Model, we have seen the encoder takes the output of the final fully connected layer of CNN and final LSTM unit state as the outputs. While these features represent the whole image and question,no specific priority is given to certain parts of the question or certain portions of the image. In Hierarchical Co-Attention Model, we consider multiple features w.r.t image and question and give priority to certain features. **The priority given to certain features is called attention.** In our model, we consider attention to image features based on question features and attention to question features based on image features. This is so called co-attention. Before explaining the model, let's first know about the attention mechanism, features considered in Hierarchy Model and co-attention mechanism.


LSTM
LSTM
LSTM
...
LSTM
LSTM question encoding
Max-over different filter pooling layer
Convolution layer with multiple filter of different widths
Word embedding
“What color on the ... up ?”

GitHub page : https://github.com/jiasenlu/HieCoAttenVQA

Reference :

1. https://github.com/Heronalps/Visual_QA_Attn
2. https://www.arxiv-vanity.com/papers/1606.00061/
3. https://medium.com/ai2-blog/may-i-have-your-attention-please-eb6cfafce938

3. **Other related Papers:**:

a. Question Type Guided Attention in Visual Question Answering

b. Stacked Attention Networks for Image Question Answering

c. Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering

d. Tips and Tricks for Visual Question Answering: Learnings from the 2017 Challenge

e. BLOG : BADRIPATRO

f. BLOG : Pythia v0.1: the Winning Entry to the VQA Challenge 2018

g. BLOG : A software suite for Visual Question Answering