

**VISUAL QUESTION ANSEWR**

**Capstone Project Report**

April – July 2019

PGP Data Science, Business Analytics and Big Data

DF1809-CM

**Team Member:**

Urvi Parekh

Manvai Agrawal

Abhinav Rahate

**Abstract**

In this report we document our experiment of building visual question answering. The experiment is, given image and open-ended question asked about the image we have to predict an answer to it. We have chosen the COCO-Image dataset where there are around 82K images & 4L question-answer pairs. For one particular image there are 4 questions attached to it. We begin with researching about the problem and done research about the state-of-art.

Due to limitation of the processing power we went ahead with 10k images and respective questions for them. In this report we investigate and document the problem of VAQ which can be used in educating a child playing a game on a touch screen.

**Acknowledgement**

“It is not possible to prepare a project without the assistance and encouragement of other people. This one is certainly no exception.”

We are extremely thankful and extend our gratitude towards AEGIS SCHOOL OF DATA SCIENCE and its faculty for their valuable guidance and support on completion of this project.

Special thanks to Jay sir for guiding us in very difficult phase of our project.

Any Omission in this brief acknowledgement does not mean lack of gratitude.

Thank you.

Urvi Parekh

Abhinav Rahate

Manavi Agrawal

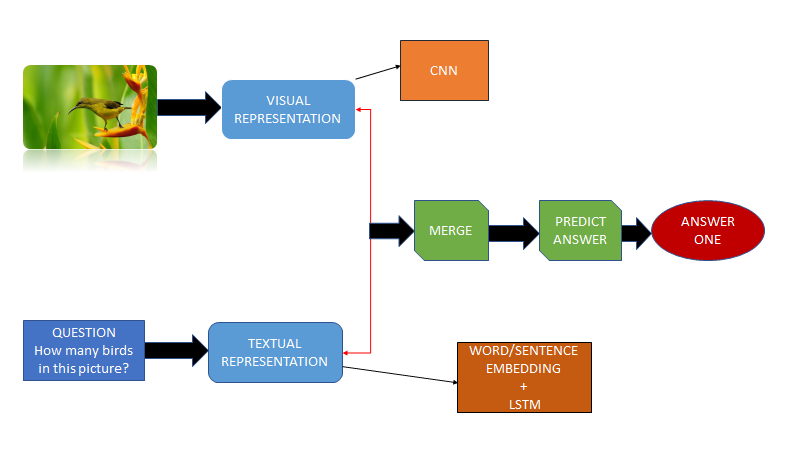
**Introduction**

The Visual question answering challenge was started by Virginia Tech in 2015 and going-on since then. The main applications of this project are:

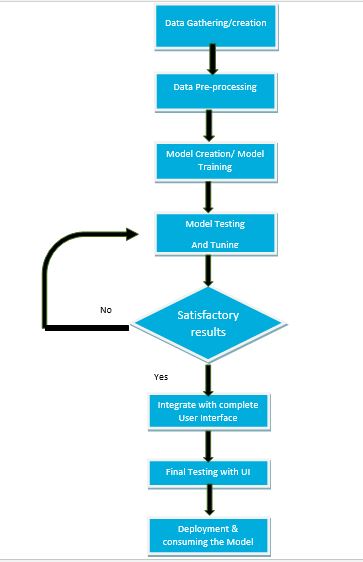
1. Educating a child playing a game on a touch screen.
2. Providing information to a spectator at an art gallery, or interacting with a robot.
3. Shopping apps like amazon and Flipkart could make their product search just so much better, like you could specify the particular type of design and it would be there

**DESIGN**

1. Import and mapping the Coco-Image dataset.
2. Pre-processing the image size and questions and answers.
3. Extracting image features with VGG16 Architecture.
4. Tokenizing and Padding the Question.
5. Calculating the Glove Weights.
6. Building image\_model and language\_model
7. Training the data by combining both the models
8. Analyzing the Accuracy



**Project Lifecycle**



# IMPLEMENTATION

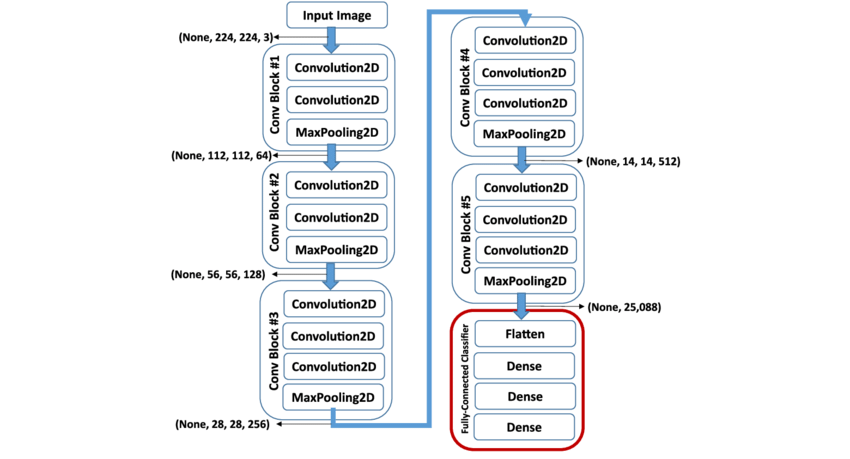
## Data Pre-Processing

1. Pre-processed questions (Using nltk).
2. Tokenization and Padding questions.
3. Calculating the glove weights.
4. Reshaping the image into 224 \*224\*3 as this is pre requisite of VGG16.
5. Extracting Image features.
6. Manipulated VGG16 architecture by removing last 2 layer(softmax) and train only last two layers of VGG16 to extract the features.
7. Converting the answer label into integer form (using LabelEncoder).

**Model Building**

VGG16:

**VGG16** (also called OxfordNet) is a convolutional neural network architecture named after the Visual Geometry Group from Oxford, who developed it. It was **used** to win the ILSVR (ImageNet) competition in 2014.

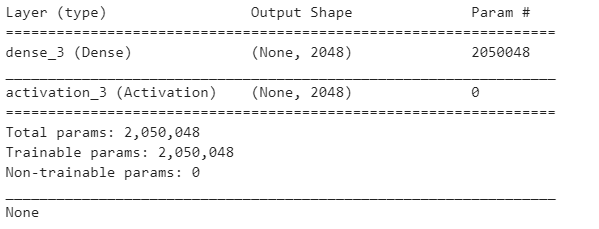


We have used VGG16 for feature extraction from an image by removing last 2 layers of VGG16 and trained only last 2 layers of the remaining model.

Why VGG16?

It is a simpler model and classifies correctly around 90% of the times amongst 1000 categories.

1.Image Model



* It is a single dense layer neural network with a shape of 2048.
* Later on in a newer version we added a dropout layer.

2.Language\_Model

For building a language model we use LSTM (Long Short Term Memory ) Deep learning or deep structured learning can be defined as special kind of neural networks composed of multiple layers.

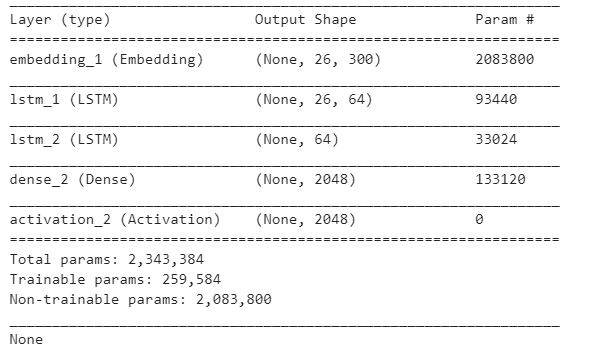
These networks are better than traditional neural network in persisting the information from previous event.

Recurrent neural network(RNN) is one such machine that has a combination of networks in loop. The networks in loop allow the information to persist. Each network in the loop takes input and information from previous network, performs the specified operation and produces output along with passing the information to next network.

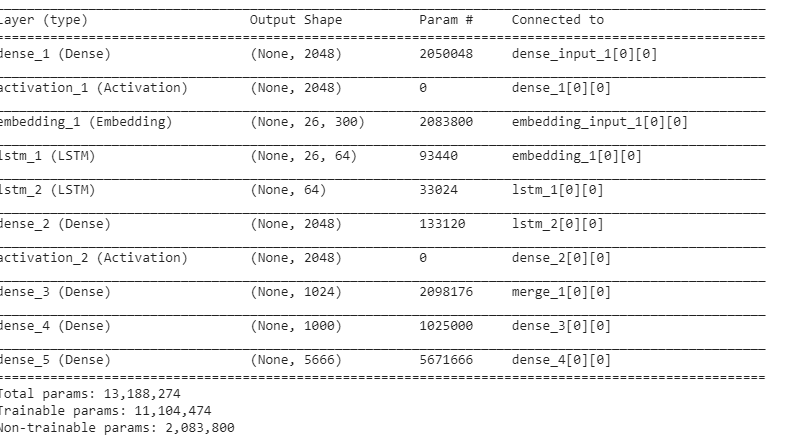
Why LSTM?

Problem of using RNN:

As more layers using certain activation functions are added to neural networks, the gradients of the loss function approaches zero, making the network hard to train this isssue is called Vanishing Gradient.

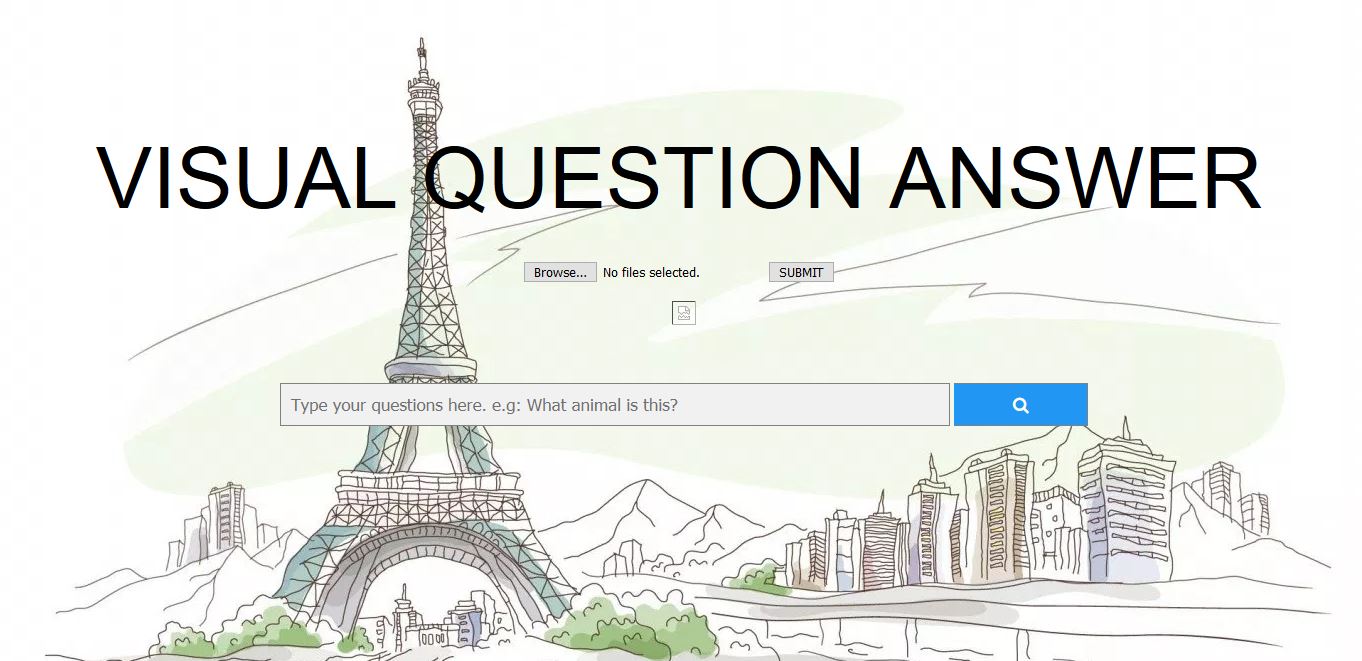


Combine Model

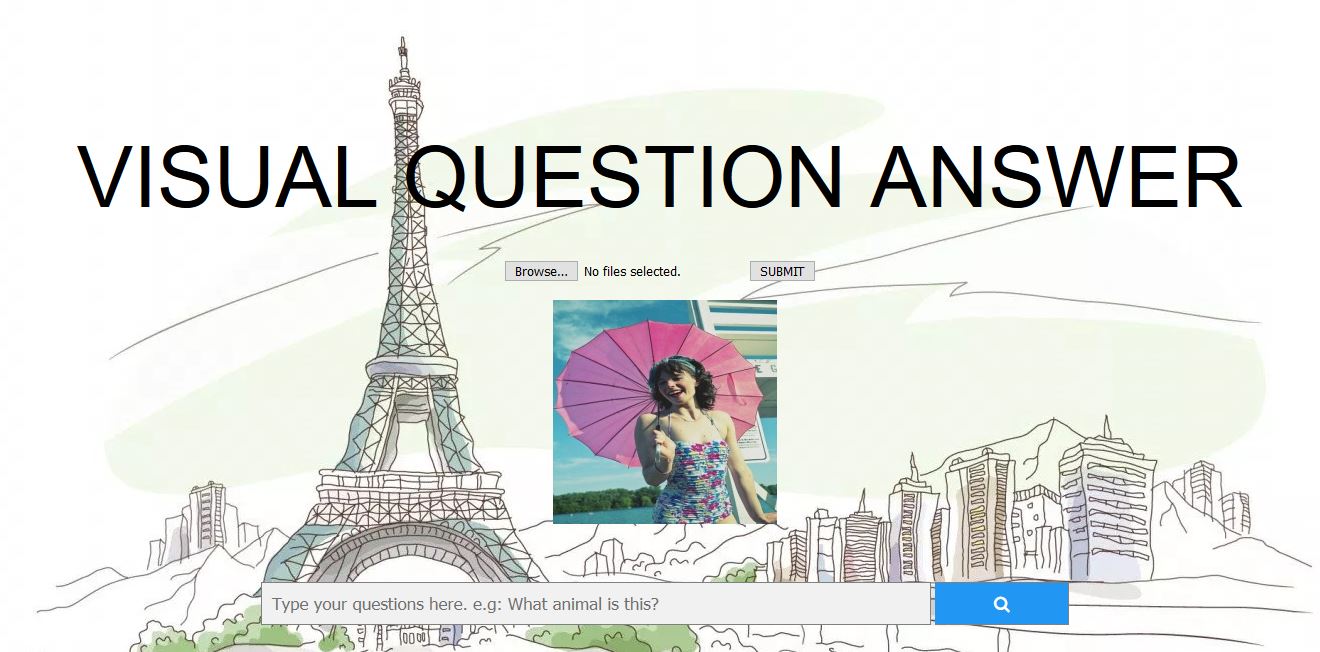


**Deployment**

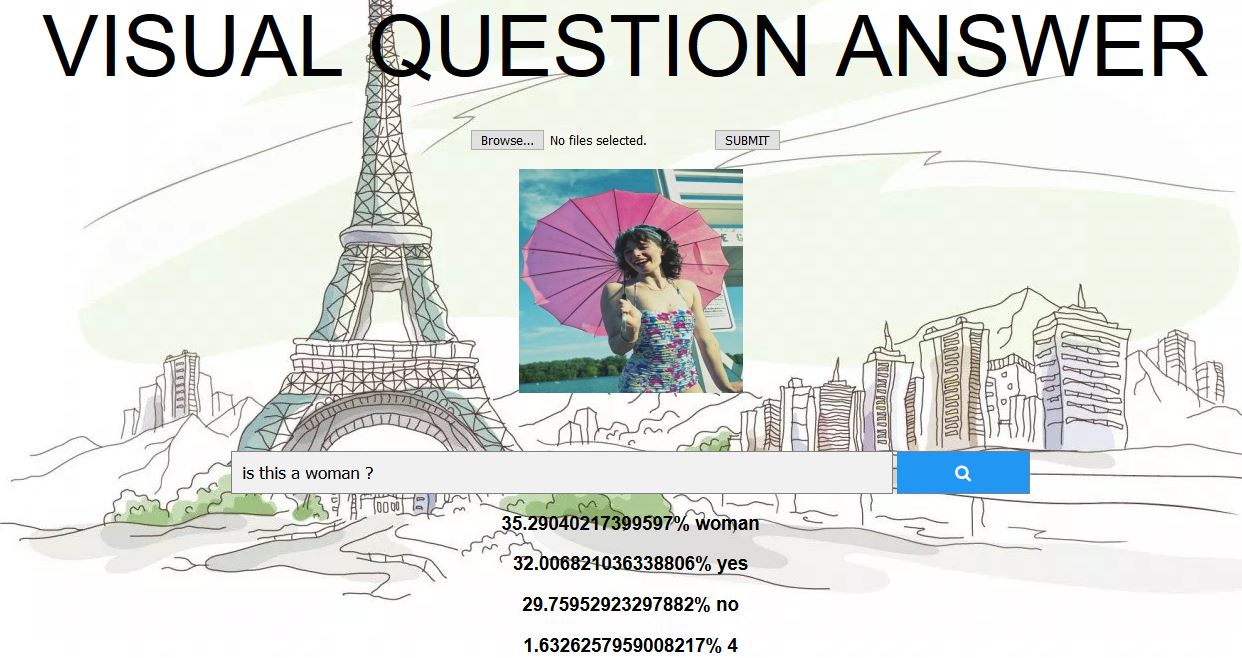
For the Deployment Purpose we use a flask wed frame work It is classified as a [microframework](https://en.wikipedia.org/wiki/Microframework) because it does not require particular tools or libraries. It has no database abstraction layer, form validation, or any other components where pre-existing third-party libraries provide common functions. However, Flask supports extensions that can add application features as if they were implemented in Flask itself. Extensions exist for object-relational mappers, form validation, upload handling, various open authentication technologies and several common framework related tools.



The above image is our front web page which is design in HTML



Here we can browse the particular image and submit that image for image processing and we can type a question which is related to image in the search bar



From the above image we can see that we are able to predict with the highest probability of woman.

**Challenges**

Various Challenges Faced:

1. To handle data of different type (questions features and image features)
2. Every row had array of 30 as question features and array of 1000 as image features. Because this we had a lot of trouble while passing it to embedding layer and dese layer respectively as it was throwing shape errors.
3. To merge two layers “Merge“ was removed in newer version of keras .
4. To combine models both model had to have same shape.
5. Issue we are still facing is processing power because of which we could only use 1000 images.

**LIMITATION**

* 1. Due to limitation of the processing power we went ahead with 10k images and respective questions for them.
  2. As this is very complicated problem and needs huge amount of data to process and make the vocabulary accordingly we were not able to reach acceptable accuracy.

**Reference Research Papers**

* <https://arxiv.org/pdf/1708.02711v1.pdf>
* <https://arxiv.org/pdf/1705.06676v1.pdf>
* <https://arxiv.org/abs/1606.00061>
* <https://www.coursehero.com/file/36229255/150500468pdf/>
* <http://openaccess.thecvf.com/content_cvpr_2017/papers/Goyal_Making_the_v_CVPR_2017_paper.pdf>
* <https://arxiv.org/abs/1612.00837>
* <http://openaccess.thecvf.com/content_cvpr_2016/papers/Shih_Where_to_Look_CVPR_2016_paper.pdf>

**Conclusion**

We Achieved 40% training and testing accuracy

**Visual Question Answering** is an interesting challenge combing different disciplines, including computer vision, natural language understanding, and deep learning.