# Visual Question Answering

Aathira Satheesh
Anagha K
Fathima Salim
Jerin Jayaraj

June 10, 2020

# Content

# Abstract

We propose the task of free-form and open-ended Visual Question Answering (VQA). Given an image and a natural language question about the image, the task is to provide an accurate natural language answer.

The VQA field is so complex that a good dataset should be large enough to capture the long range of possibilities within questions and image content in real world scenarios.

Visual questions selectively target different areas of an image, including background details and underlying context. As a result, a system that succeeds at VQA typically needs a more detailed understanding of the image and complex reasoning.

# Introduction

- In our VQA model we are using pretrained weights that is obtained from training imagenet images on VGG19 model.
- Glove vectors are used to convert words into vector form by aggregating global word-word co-occurrence matrix from a corpus.
- VGG19 is a convolutional neural network model which achieves 90% top-5 test accuracy in ImageNet.
- Keras has been used for implementing the neural network

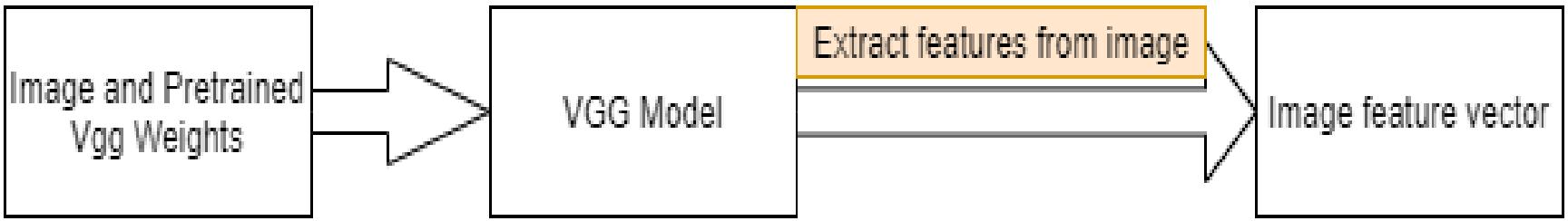
# Problem Statement

On seeing an image when a question is asked , answering it is a challenging task . As a human we can answer this pretty easily but for a machine to answer it , the machine needs to learn a lot of things . The search and the reasoning part must be performed over the content of an image.

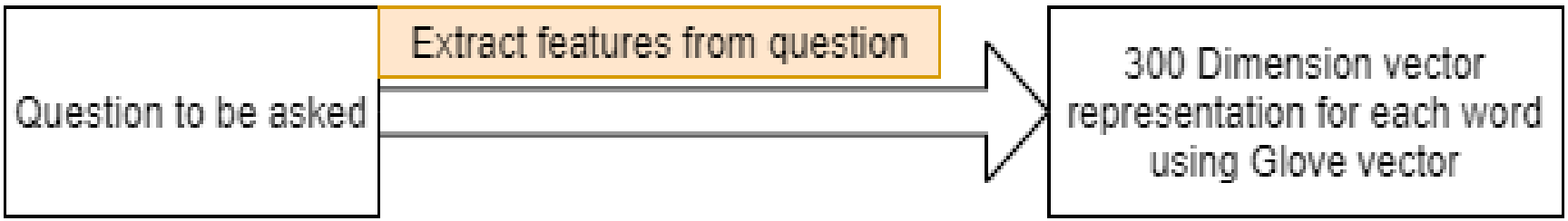
We need to create a model that predicts the answer of an open-ended question related to a given image.

# Block diagrams

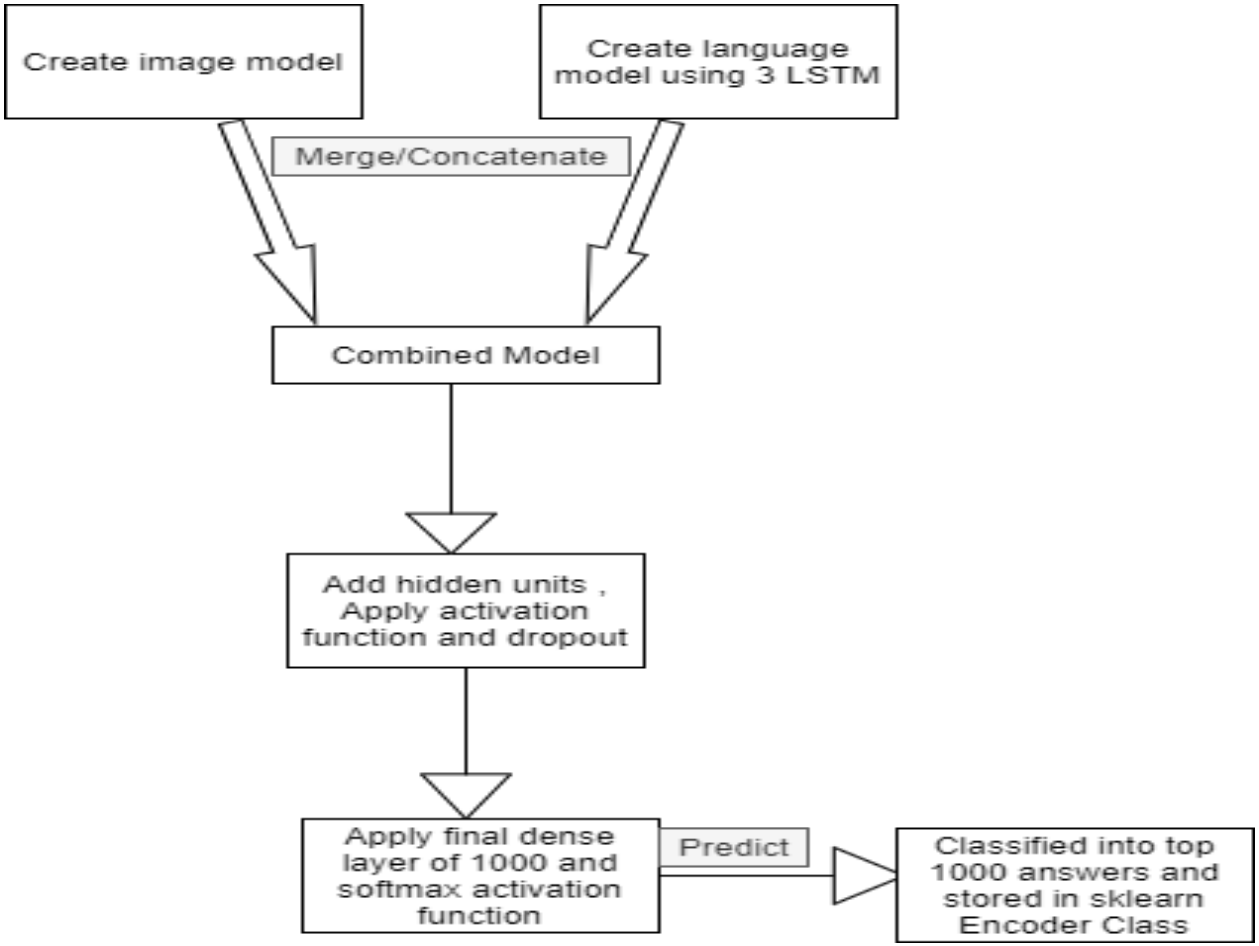
## 1)Extracting image features




## 2)Extracting question feature



## 3) Combined VQA Model



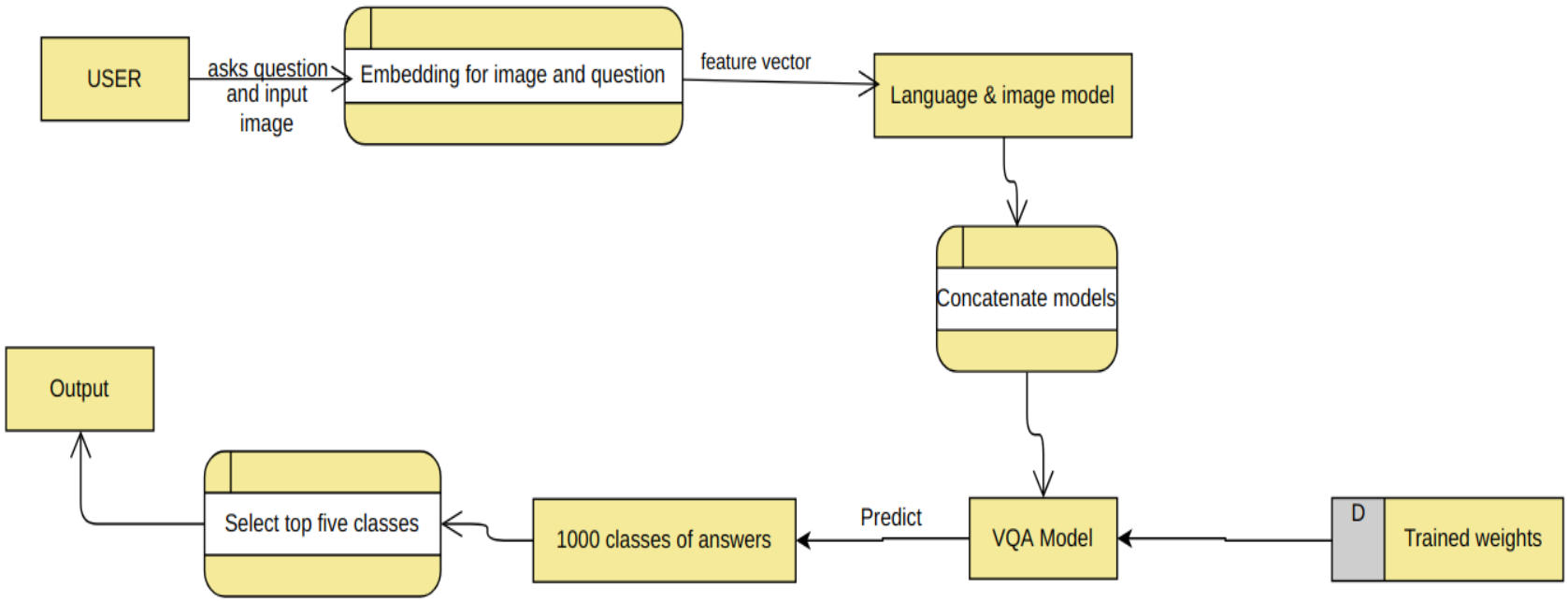
# Algorithm

- First input an image
- Image is resized into 224*224 (Since VGG was trained on images of this size)
- Image features are extracted by giving image file to VGG model and using the pre-trained weights
- Next input the question we want to ask
- Each word of the question is transformed to a 300 dimensional representation produced by spaCy.spaCy uses an algorithm called GloVe which reduces a given token into 300 dimensional representation.
- These features are then passed to a VQA Model which is a combination of LSTM and MLP .
- This VQA model was trained on COCO train set available on VQA website for questions.

# Algorithm Continued

- The images for VQA model were trained on VGG19 model.
- Our VQA model consists of an image model that defines structure of input image , language model that contains LSTM to memorise sequential data and then these two models are concatenated.
- Then we add dense layer , a tanh non linear activation function and also dropout for each of the three hidden layer.
- We then classify our answers to top 1000 answers.
- Finally we apply softmax layer that normalise the values to fit between 0 and 1 . It gives the probability value .

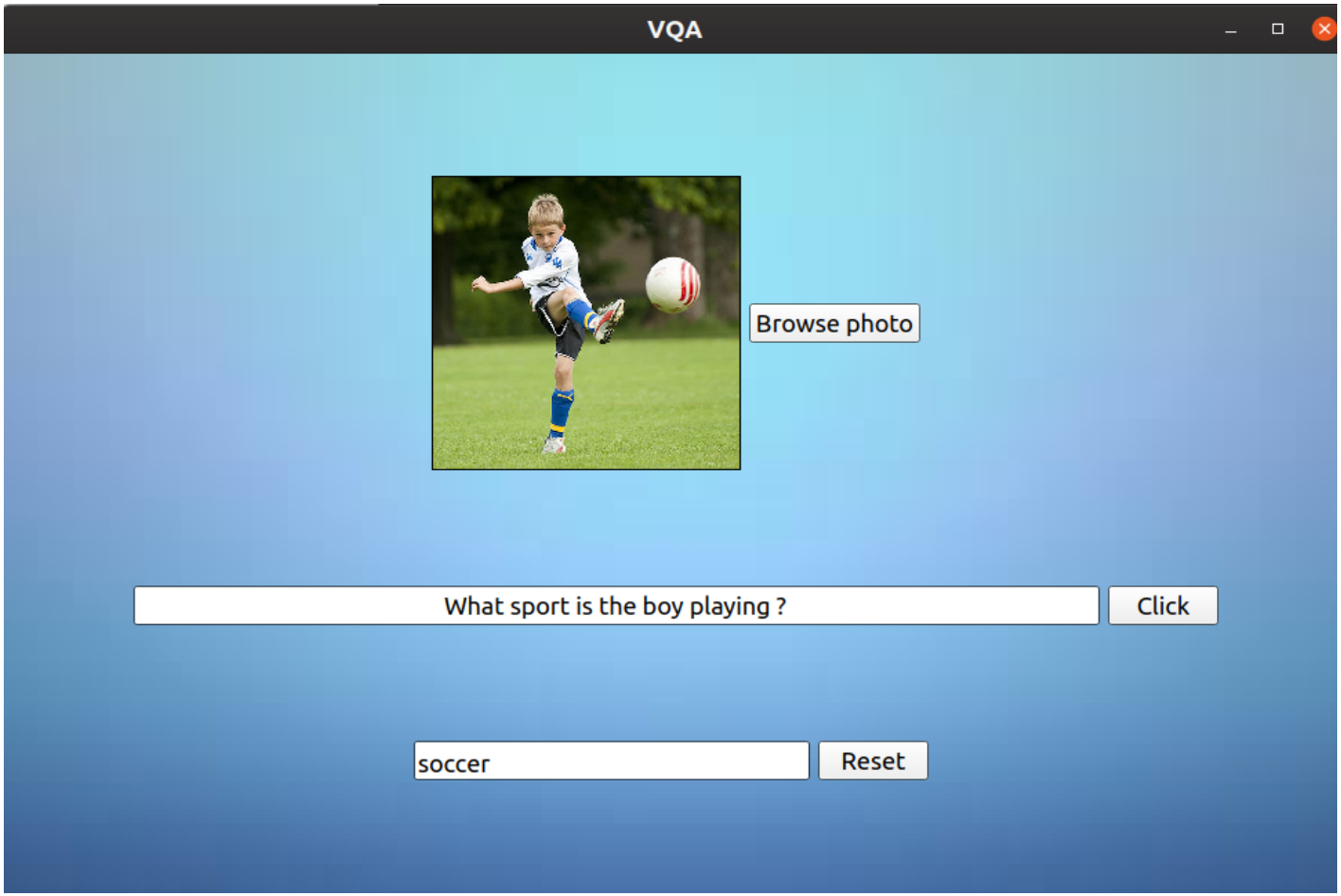
# Data Flow Diagram



# Implementation

- In order to create neural networks , we have used the Keras library in python .
- Keras has all inbuilt function to create sequential model , apply activation function , add dense layers etc
- We have implemented the project both in terminal and as a desktop application(using Qt with c++)
- On terminal based method , we passed the image and question as arguments .
- We have used a sklearn labelencoder class that contains the list of all the top 1000 classes along with their labels.
- After predicting , we get probabilities in the end since we have used Softmax layer.
- The highest probability denotes the class with the highest chance to be our answer .

# Results




On giving and image and a question , the answer is printed . Reset will reset the screen to its initial state.

# Project relevance

If this project is trained on an even more larger dataset , we can get more accurate results . the most direct application is to help blind and visually-impaired users. VQA system could provide information about an image on the Web or any social media. Another obvious application is to integrate VQA into image retrieval systems. This could have a huge impact on social media or e-commerce.

# Conclusion

- VQA is 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. In general, we can outline the approaches in VQA as follows:
  1)Extract features from the question.
  2)Extract features from the image.
  3)Combine the features to generate an answer.
- For text features, Long Short Term Memory (LSTM) encoders are used.
- In the case of image features, CNNs pre-trained on ImageNet(VGG19) is the most frequent choice.
- Regarding the generation of the answer, the approaches usually model the problem as a classification task.

# References

[1]. www.keras.io - Keras Documentation
[2]. ieeexplore.ieee.org/document/7410636
[3]. github.com/iamaaditya/VQADemo
[4].tryolabs.com/blog/2018/03/01/introduction-to-visual-question-answering/
[5]. github.com/anujshah1003/owndatacnnimplementationkeras.git