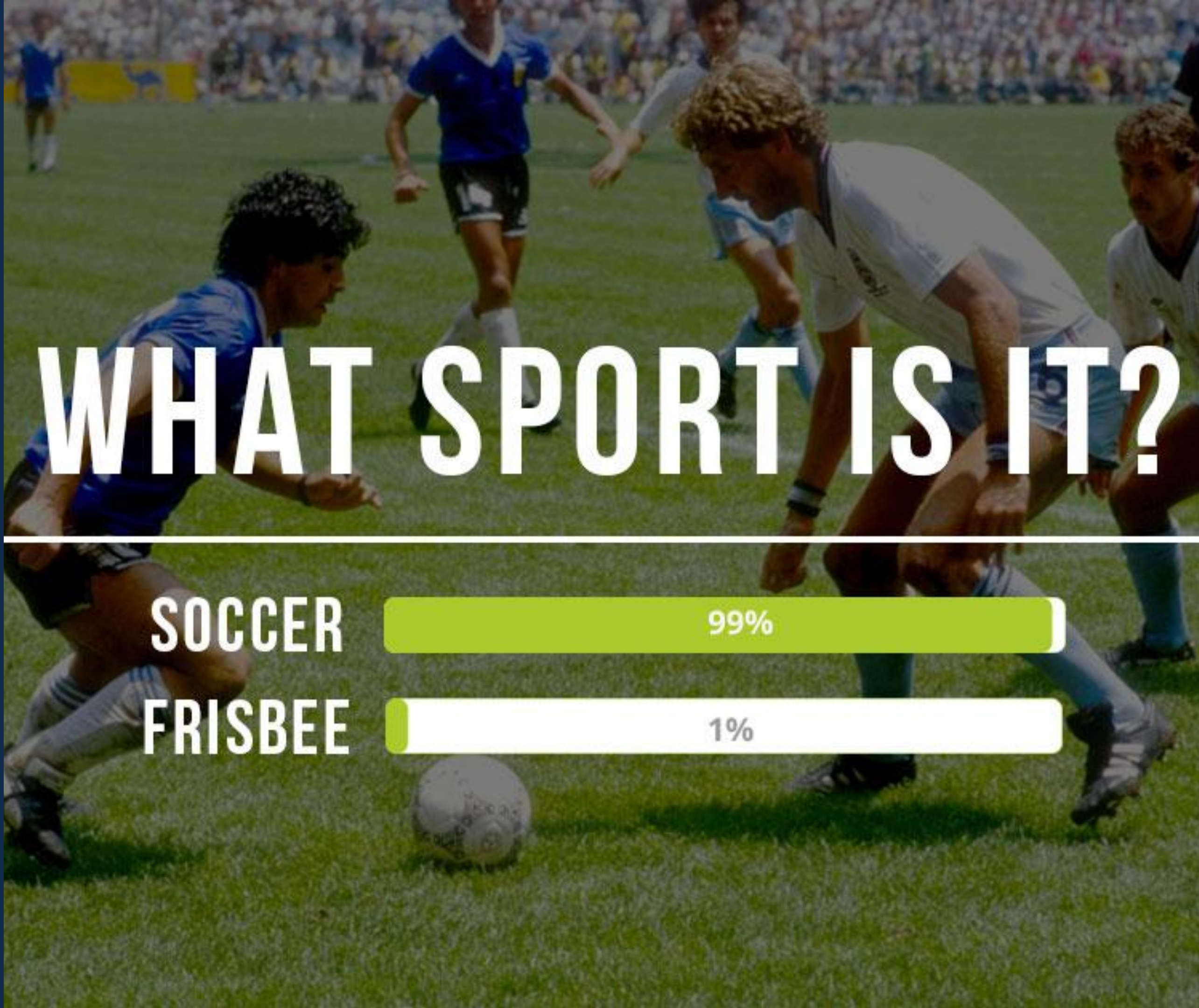
Visual Question Answering

Presented By :

Urvi Parekh

Abhinav Rahate

Manavi Agrawal



Contents

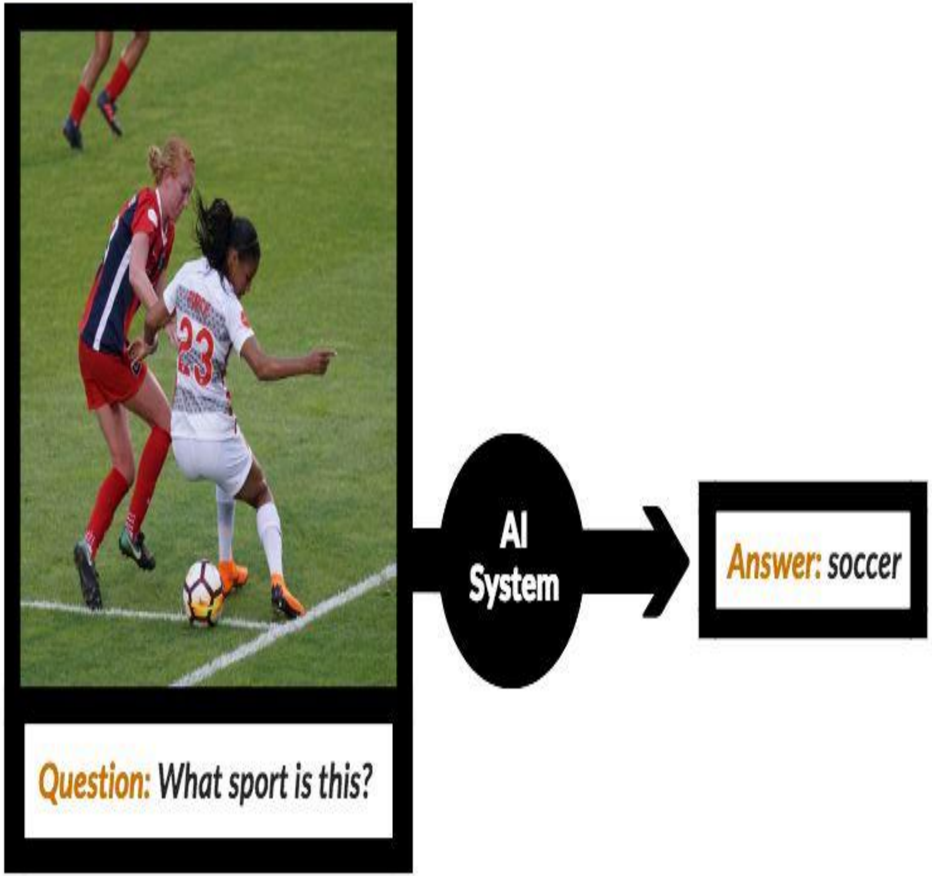
1. What is VQA?
2. Applications
3. Workflow
4. Data Pre-Processed
5. Model
6. Deployment
7. Challenges
8. Research Methodology
9. Limitation

Aegis
SCHOOL OF DATA SCIENCE

What is VQA?

Given an image, can our machine answer the corresponding questions in natural language?

What is VQA?



What is VQA?

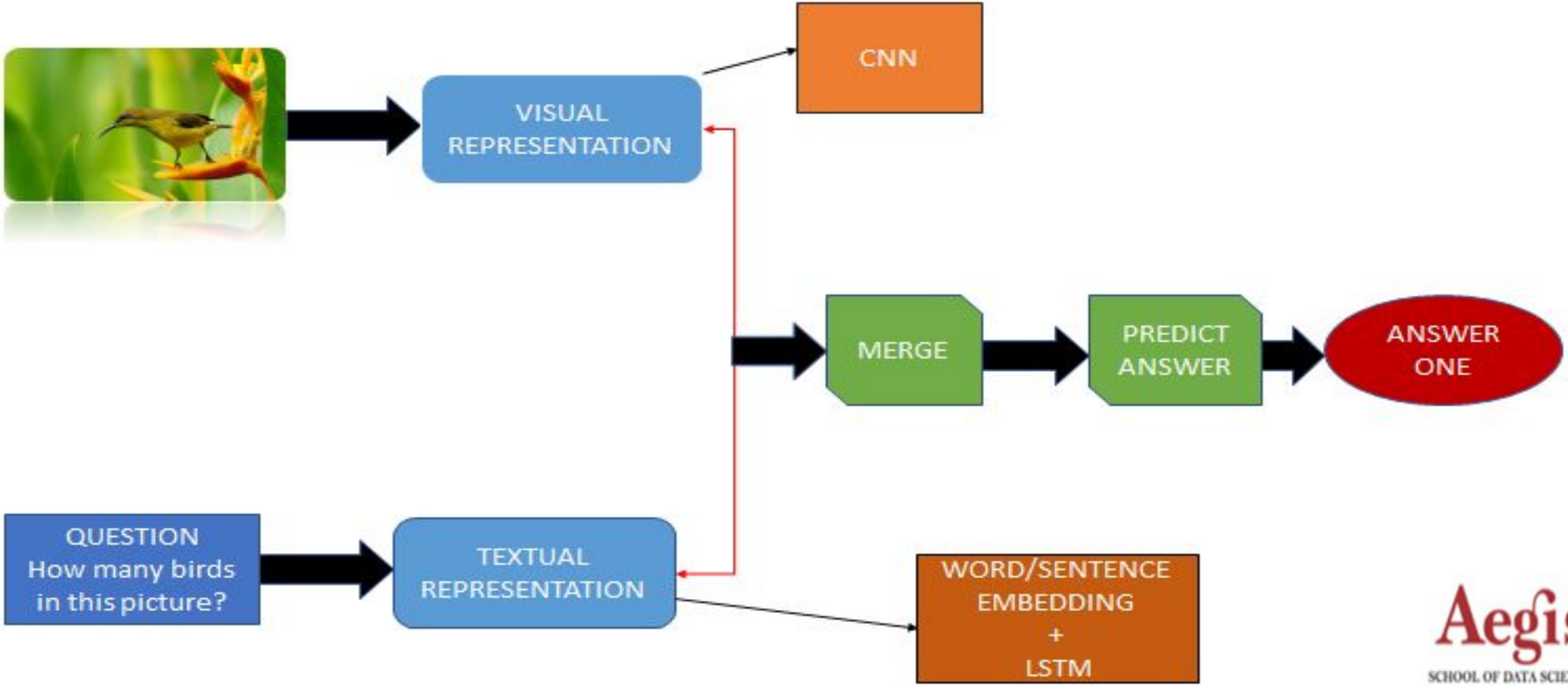
In order to do this, our model would need to understand several things - let’s break them down into sub-tasks:

1. Identifying the various objects in the image (the train, traffic signals, tracks, pavement, person, etc)
2. Processing the text of the question itself, which can be processed as a ‘sequence’ of words
3. Mapping the appropriate sections of the image (in this case - the train) to the input text question.
4. Generating natural language text in the form of an answer.

Applications

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.

Workflow



CNN

VISUAL REPRESENTATION

MERGE

PREDICT ANSWER

ANSWER ONE

QUESTION
How many birds
in this picture?

TEXTUAL REPRESENTATION

WORD/SENTENCE EMBEDDING
+
LSTM

Aegis
SCHOOL OF DATA SCIENCE

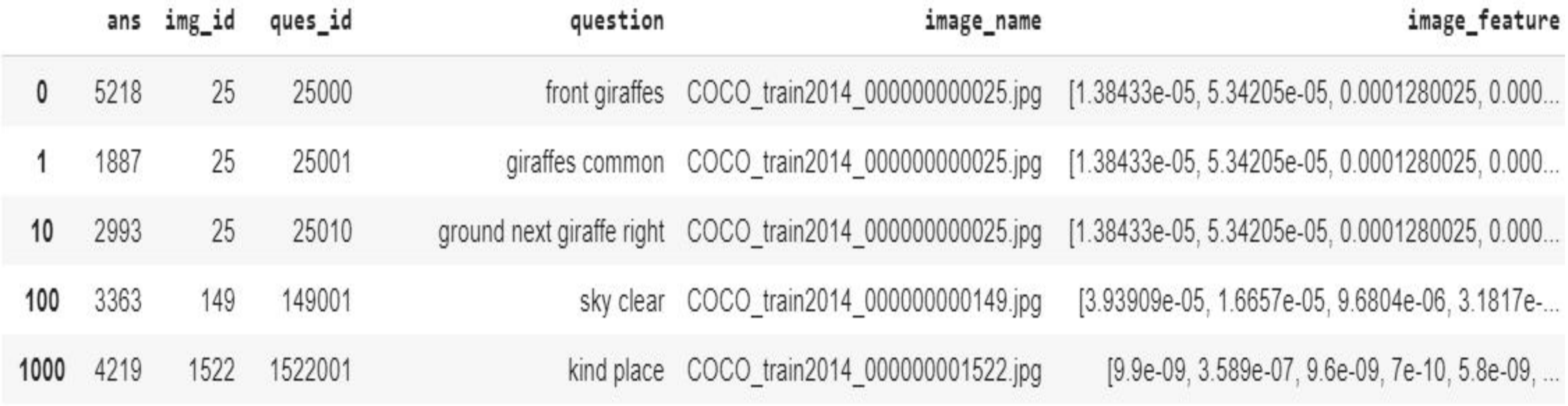
Workflow

1. Pre-Processed the questions
2. Extracted Image Features
3. Calculated Glove Weights
4. Building Image Model
5. Building Language Model
6. Combining The Model
7. Testing the Model
8. Deployment

Data Pre-Processing

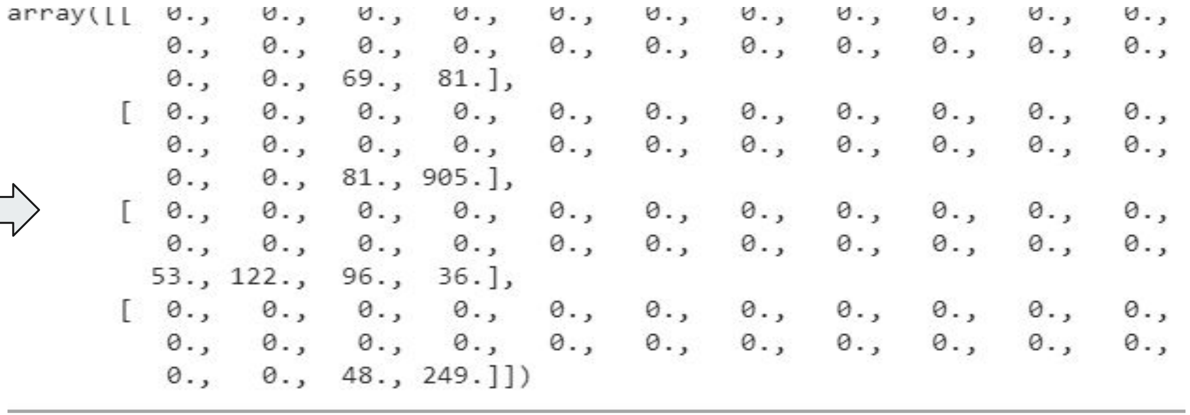
1. Pre-processed questions.(Using nltk)
2. Tokenization and Padding questions.
3. Calculating the glove weights.
4. Reshaping the image into 224 *224*3
5. Extracting Image features.(using pre-trained model VGG16)
6. Converting the answer label into integer form.(using LabelEncoder)

Data Pre-Processing



	ans	img_id	ques_id	question	image_name	image_feature
0	5218	25	25000	front giraffes	COCO_train2014_000000000025.jpg	[1.38433e-05, 5.34205e-05, 0.0001280025, 0.000...
1	1887	25	25001	giraffes common	COCO_train2014_000000000025.jpg	[1.38433e-05, 5.34205e-05, 0.0001280025, 0.000...
10	2993	25	25010	ground next giraffe right	COCO_train2014_000000000025.jpg	[1.38433e-05, 5.34205e-05, 0.0001280025, 0.000...
100	3363	149	149001	sky clear	COCO_train2014_000000000149.jpg	[3.93909e-05, 1.6657e-05, 9.6804e-06, 3.1817e-...
1000	4219	1522	1522001	kind place	COCO_train2014_000000001522.jpg	[9.9e-09, 3.589e-07, 9.6e-09, 7e-10, 5.8e-09, ...

Question features

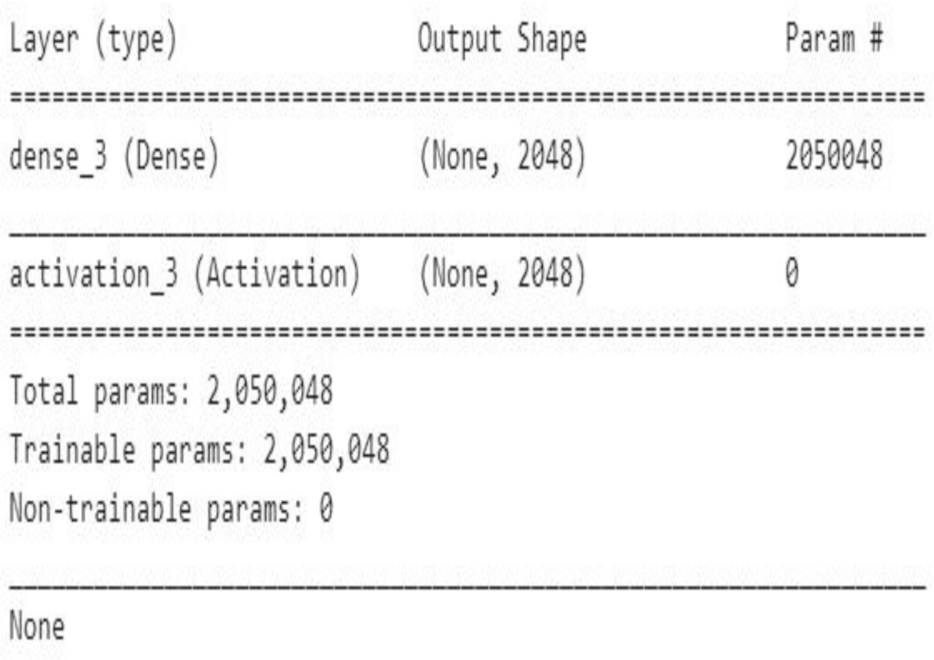


```
array([[  0.,   0.,   0.,   0.,   0.,   0.,   0.,   0.,   0.,   0.,   0.,
          0.,   0.,   0.,   0.,   0.,   0.,   0.,   0.,   0.,   0.,   0.,
          0.,   0.,  69.,  81.],
       [  0.,   0.,   0.,   0.,   0.,   0.,   0.,   0.,   0.,   0.,   0.,
          0.,   0.,   0.,   0.,   0.,   0.,   0.,   0.,   0.,   0.,   0.,
          0.,   0.,  81., 905.],
       [  0.,   0.,   0.,   0.,   0.,   0.,   0.,   0.,   0.,   0.,   0.,
          0.,   0.,   0.,   0.,   0.,   0.,   0.,   0.,   0.,   0.,   0.,
         53., 122.,  96.,  36.],
       [  0.,   0.,   0.,   0.,   0.,   0.,   0.,   0.,   0.,   0.,   0.,
          0.,   0.,   0.,   0.,   0.,   0.,   0.,   0.,   0.,   0.,   0.,
          0.,   0.,  48., 249.]])
```

Model

Steps followed:

1. Image Model



```
Layer (type)                 Output Shape              Param #
=================================================================
dense_3 (Dense)              (None, 2048)              2050048
_________________________________________________________________
activation_3 (Activation)    (None, 2048)              0
=================================================================
Total params: 2,050,048
Trainable params: 2,050,048
Non-trainable params: 0
_________________________________________________________________
None
```

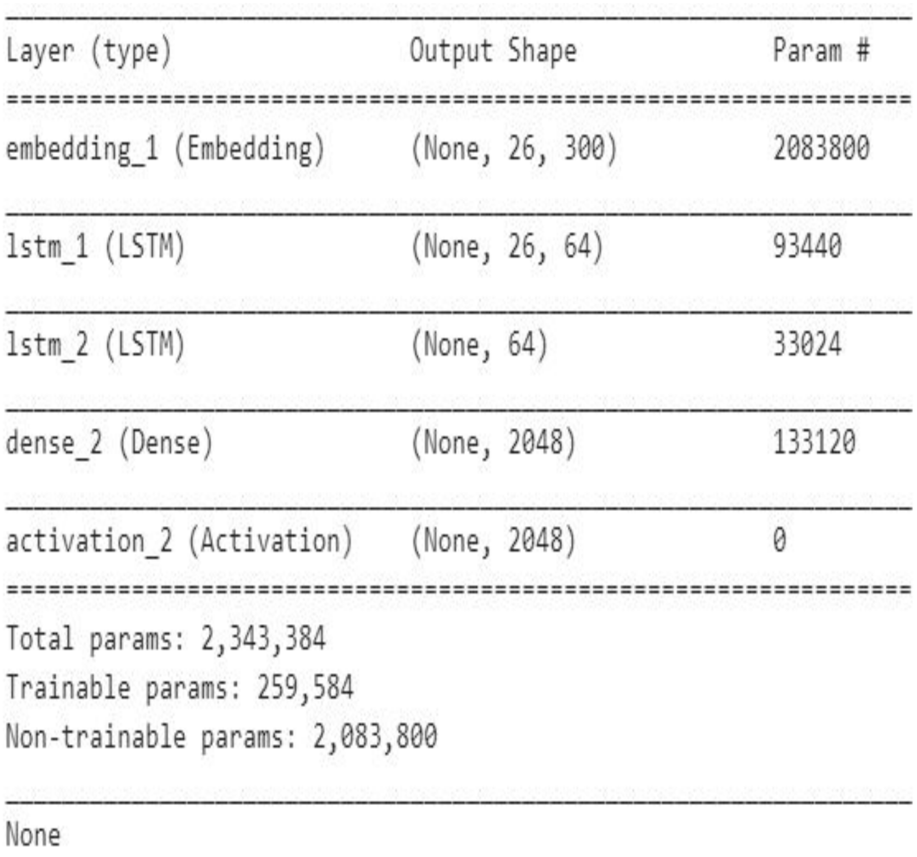
Model

2. Language Model

1. The language model is build using LSTM, which is recurrent neural network.
2. Problem of using RNN:
   a. Vanishing Gradient Descent
      i. 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.

Model

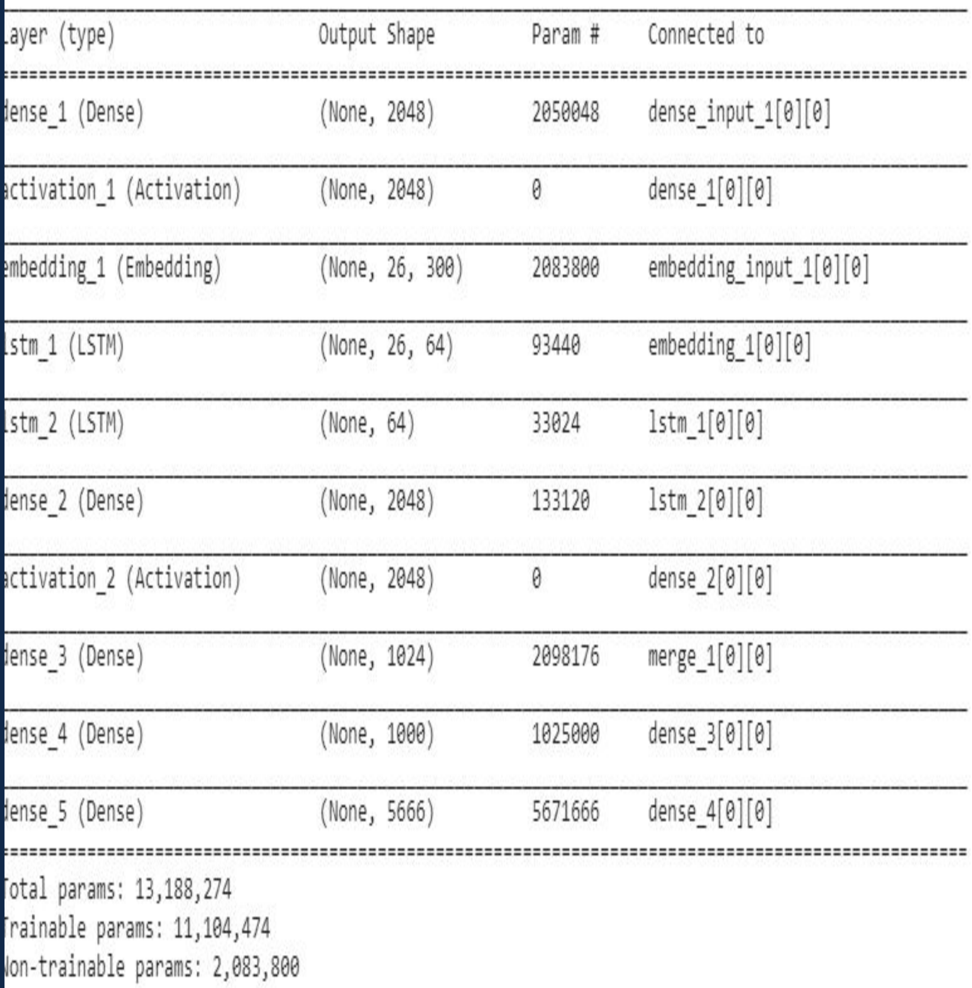
2. Language Model



```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
embedding_1 (Embedding)      (None, 26, 300)           2083800
_________________________________________________________________
lstm_1 (LSTM)                (None, 26, 64)            93440
_________________________________________________________________
lstm_2 (LSTM)                (None, 64)                33024
_________________________________________________________________
dense_2 (Dense)              (None, 2048)              133120
_________________________________________________________________
activation_2 (Activation)    (None, 2048)              0
=================================================================
Total params: 2,343,384
Trainable params: 259,584
Non-trainable params: 2,083,800
_________________________________________________________________
None
```

Model

3. Combine Model



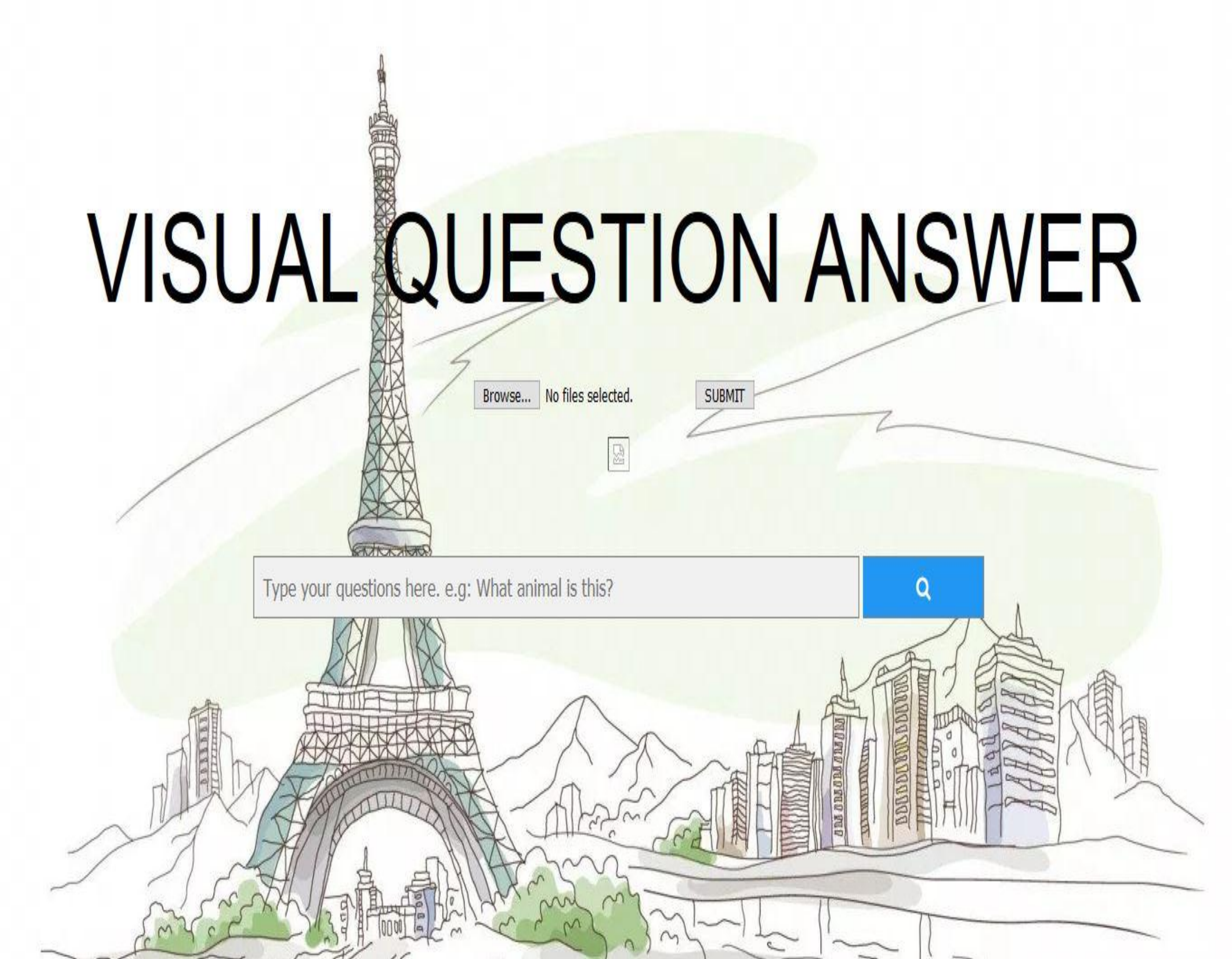
```
____________________________________________________________________________________________________
ayer (type)                      Output Shape          Param #     Connected to
====================================================================================================
lense_1 (Dense)                  (None, 2048)          2050048     dense_input_1[0][0]
____________________________________________________________________________________________________
ctivation_1 (Activation)         (None, 2048)          0           dense_1[0][0]
____________________________________________________________________________________________________
mbedding_1 (Embedding)           (None, 26, 300)       2083800     embedding_input_1[0][0]
____________________________________________________________________________________________________
lstm_1 (LSTM)                    (None, 26, 64)        93440       embedding_1[0][0]
____________________________________________________________________________________________________
lstm_2 (LSTM)                    (None, 64)            33024       lstm_1[0][0]
____________________________________________________________________________________________________
lense_2 (Dense)                  (None, 2048)          133120      lstm_2[0][0]
____________________________________________________________________________________________________
ctivation_2 (Activation)         (None, 2048)          0           dense_2[0][0]
____________________________________________________________________________________________________
lense_3 (Dense)                  (None, 1024)          2098176     merge_1[0][0]
____________________________________________________________________________________________________
lense_4 (Dense)                  (None, 1000)          1025000     dense_3[0][0]
____________________________________________________________________________________________________
lense_5 (Dense)                  (None, 5666)          5671666     dense_4[0][0]
====================================================================================================
otal params: 13,188,274
rainable params: 11,104,474
lon-trainable params: 2,083,800
```

Model Accuracy

With 10,000 images trained we are able to achieve **40% training accuracy** and **40% testing accuracy**

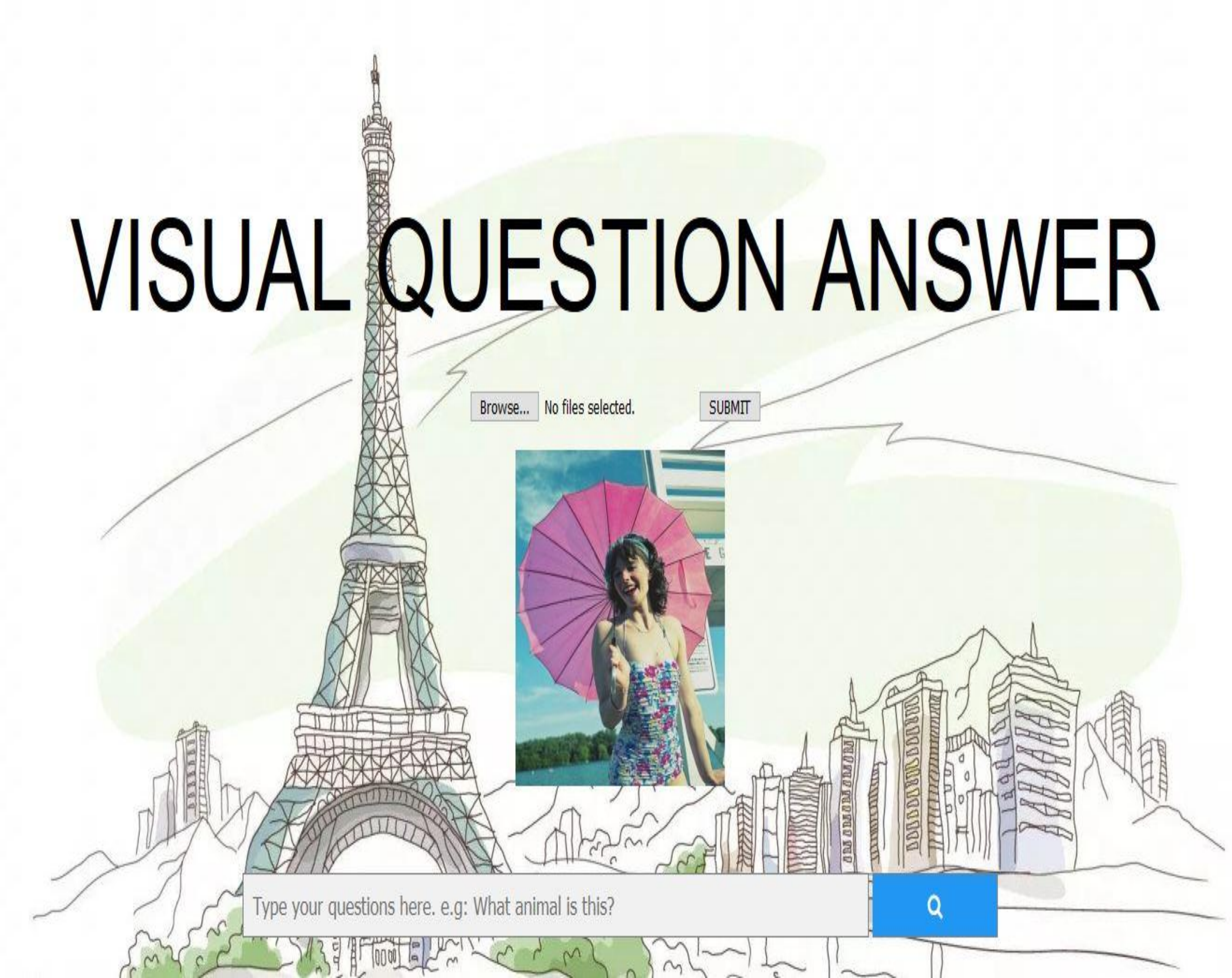
Aegis
SCHOOL OF DATA SCIENCE

Deployment



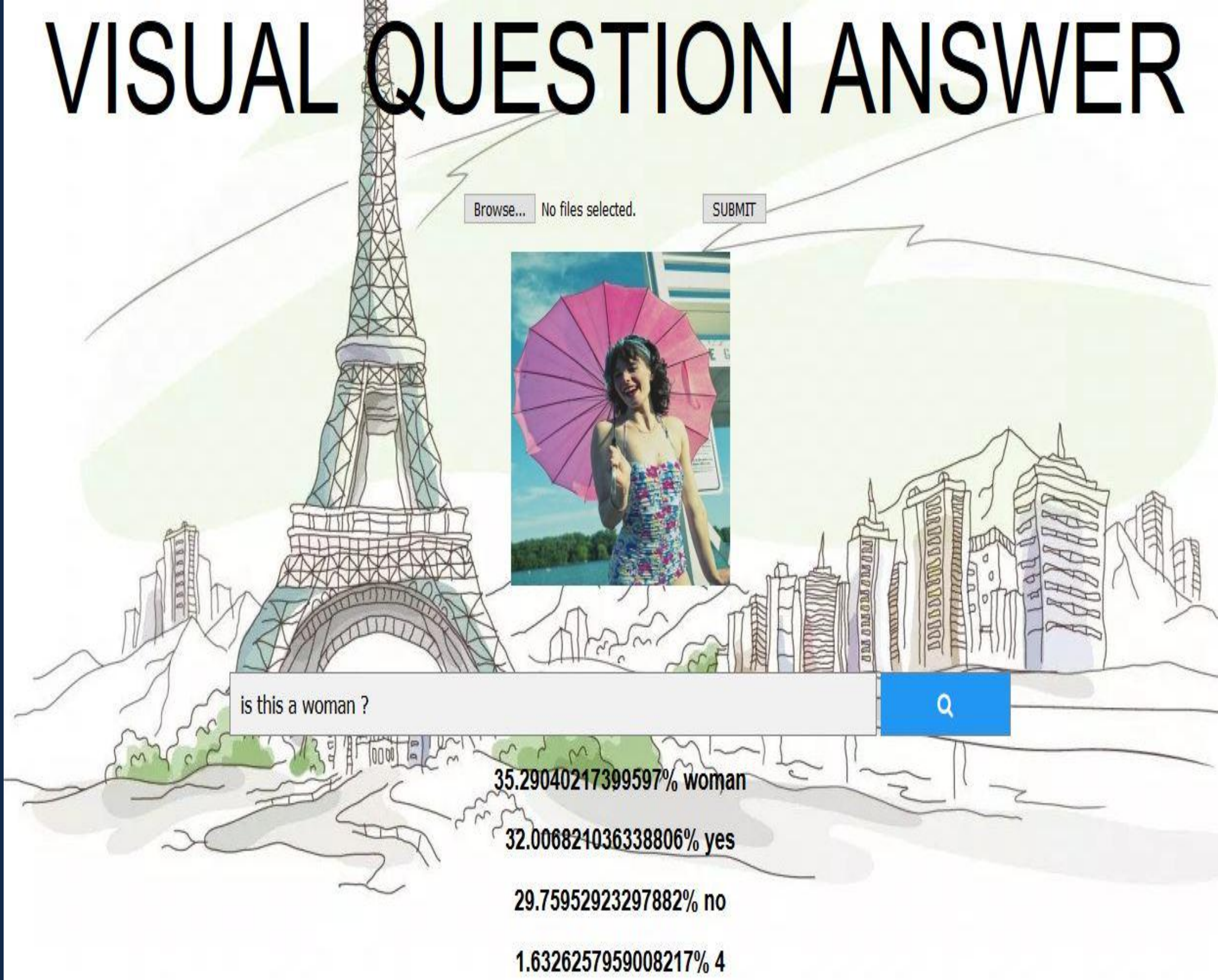
VISUAL QUESTION ANSWER

Deployment



VISUAL QUESTION ANSWER

Deployment



VISUAL QUESTION ANSWER

is this a woman ?

```
35.29040217399597% woman

32.006821036338806% yes

29.75952923297882% no

1.6326257959008217% 4
```

Challenges

1. To handle data of different type(questions features and image features)
2. Every row had array of 30 as qus 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

Aegis
SCHOOL OF DATA SCIENCE

Research Methodology

1. There has been lot of research done on VQA from last couple of years .

2. There are research papers available with different methodology for achieving the good accuracy in answering the question pertaining to the image.

3. We had taken reference from those research and methodology and build a model/product that answer more accurately .

Aegis

SCHOOL OF DATA SCIENCE

Research Methodology

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

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.

Aegis
SCHOOL OF DATA SCIENCE

THANK YOU