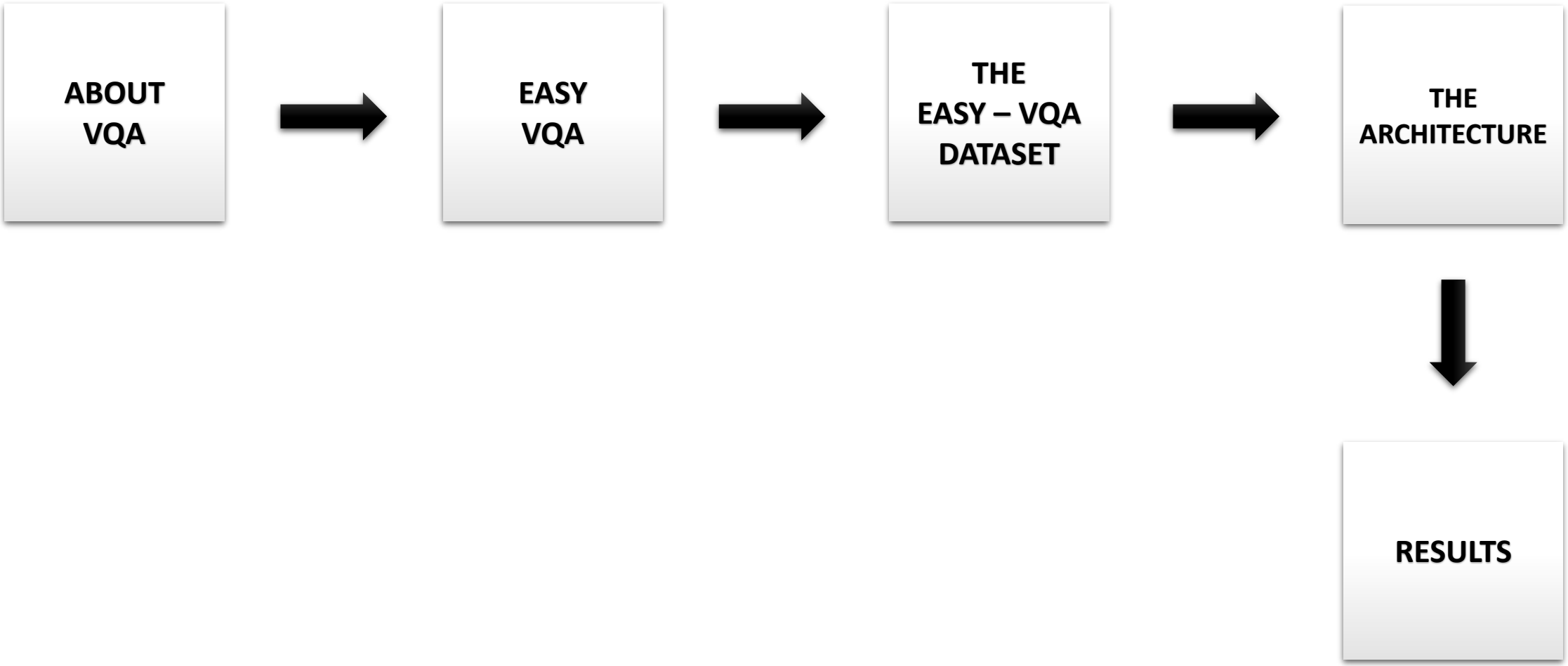
EASY
VISUAL QUESTION ANSWERING

- SAWAN AICH -

THE ROADMAP

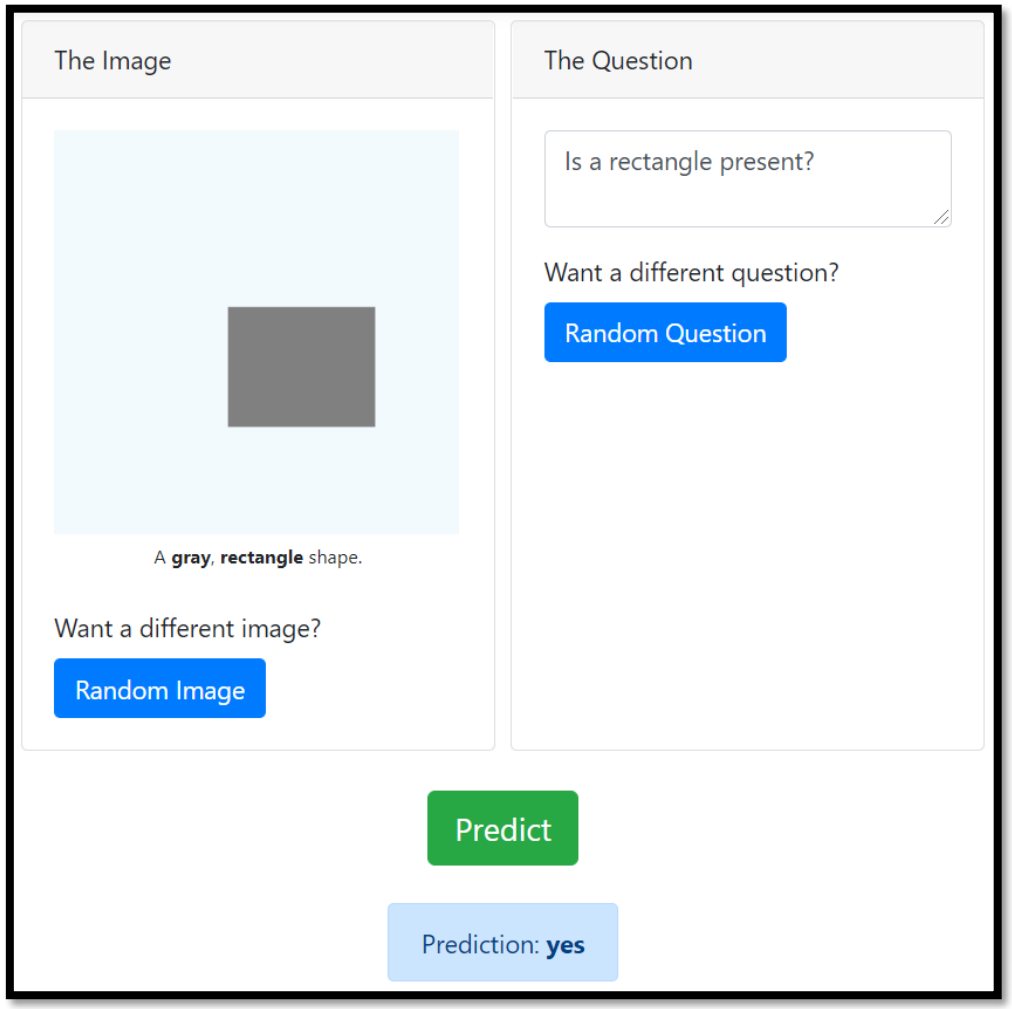

ABOUT VQA
EASY VQA
THE EASY – VQA DATASET
THE ARCHITECTURE
RESULTS

ABOUT VQA

Visual Question Answering or **VQA** is a research area about building a computer system to answer open-ended questions about an image given as input. It involves both the domains of **Computer Vision** and **Natural Language Processing**.

Applications of VQA include helping the **blind** and **visually-impaired users** and providing **information about an image** on the **Web** or any **social media**. We can also integrate VQA into **image retrieval systems**. VQA can also be used with **educational** or **recreational** purposes.

EASY VISUAL QUESTION ANSWERING



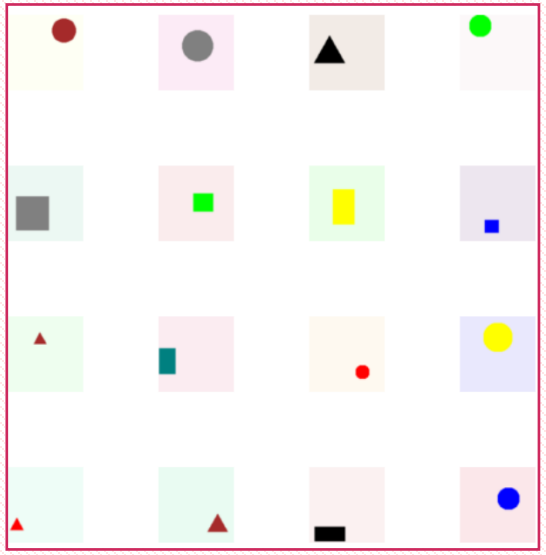
My area of interest is to work on a lighter version of VQA, namely, **Easy-VQA**. This is a demonstration of VQA on a custom dataset that is originally being created by **Victor Zhou**.

The dataset contains **5000 images** and **48,248 overall questions** with their respective answers. We split them into train, validation and test sets as per our requirements.

Applications of **Easy-VQA** include **identification of geometric shapes** in various images. Young students can find activities related to Easy-VQA quite interesting. Anyone can contribute to the dataset as it is simple and easy to implement.

EASY – VQA DATASET

IMAGES



Each image consists of either a **circle**, or a **triangle** or a **rectangle** of **8 different colors**. Each image is of size **64 x 64 x 3**. Out of **5000 images**, **3000** are taken for **training**, **1000** for **validation** and **1000** for **testing**.

QUESTIONS

"what is the blue shape?“, "does the image contain a green shape?“, "is there a rectangle?“, "what is the color of the shape”, "what is the color of the triangle?“, "does the image not contain a teal shape?“, "is no yellow shape present?"

Out of **48,248 questions**, **28,833** are taken for **training**, **9742** for **validation** and **9673** for **testing**. Each question is a combination of words taken from **a vocabulary of 27 words**.

ANSWERS

“triangle", “yes“, “no“, “gray“, “circle“, “red“, “blue“, “yellow”, “teal”, “black”, “green”, “brown”, “rectangle”

Each question has their corresponding answer to it. The answer to any question can be any of the following **13 answers** stated above. Majority of the answers to the questions are either **yes** or **no** (**35,543 answers**).

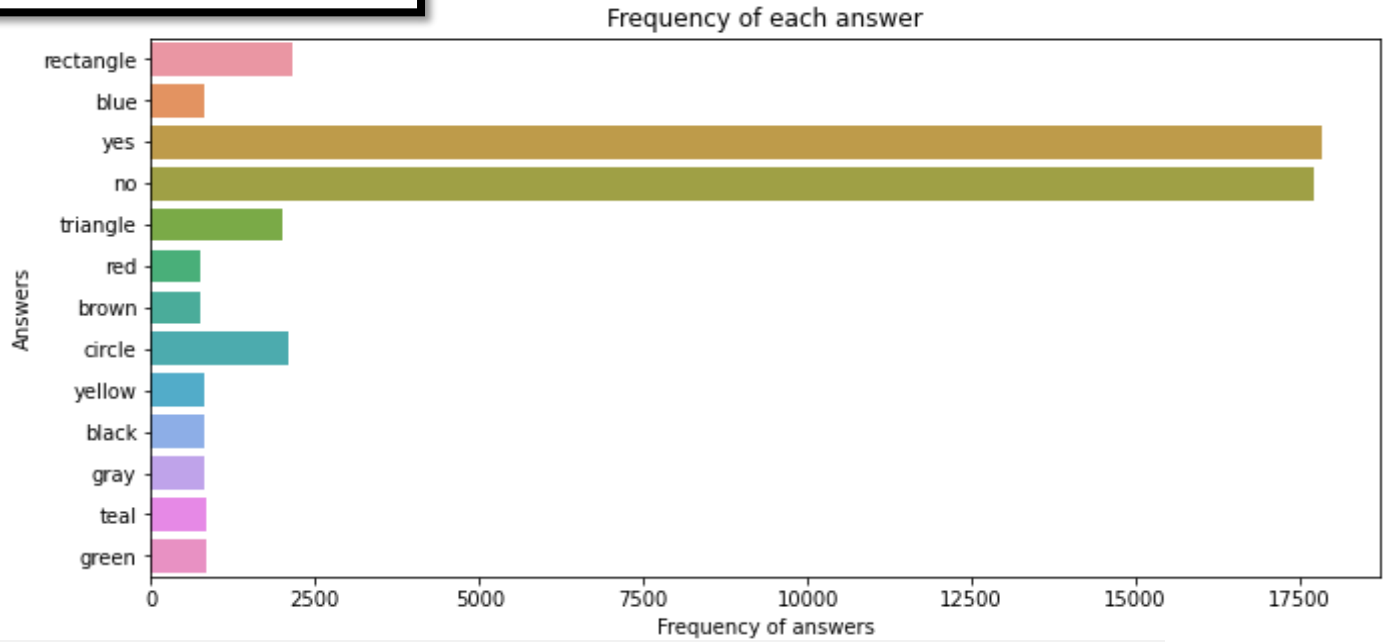
Link to Github Repository : https://github.com/vzhou842/easy-VQA

EXPLORATORY DATA ANALYSIS

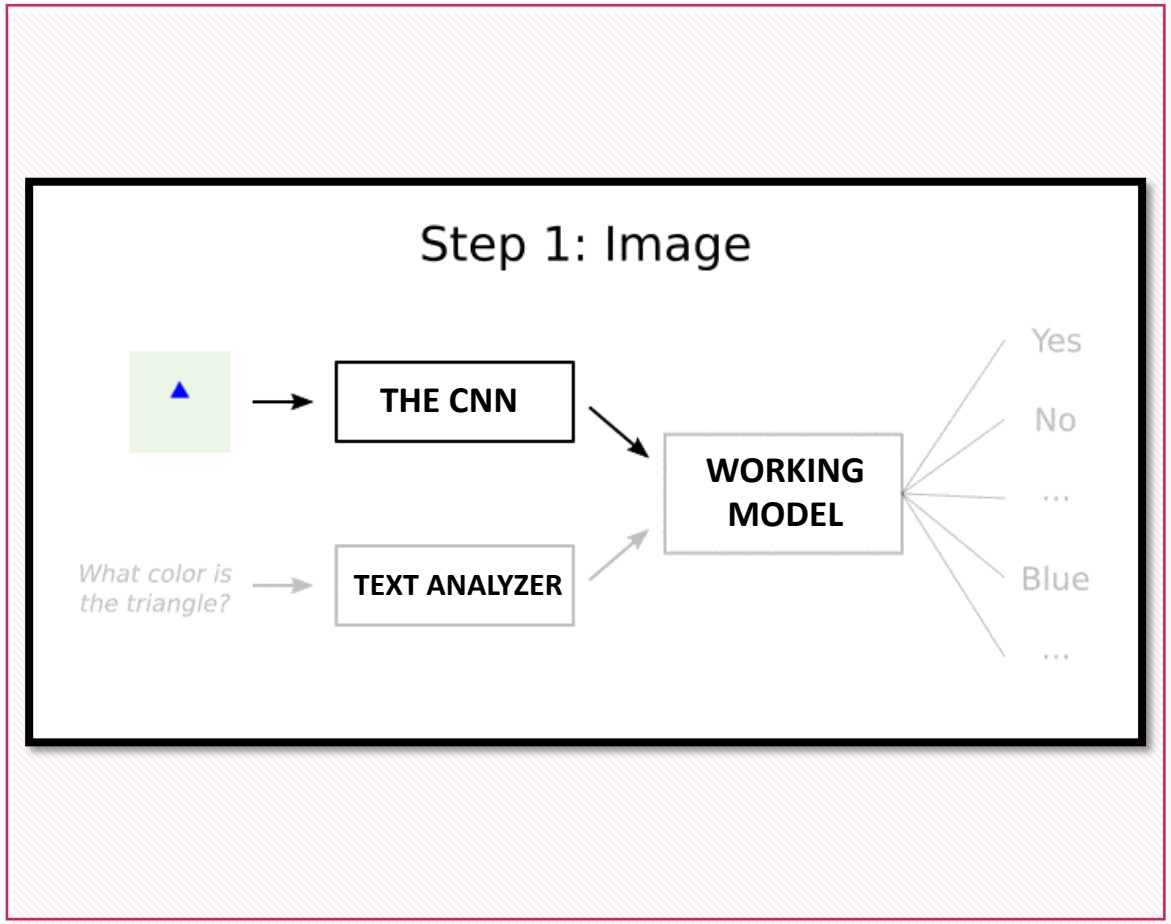
```
--- Reading questions...
Read 28833 training questions, 9742 validation questions and 9673 testing questions.

--- Reading answers...
Found 13 total answers:
['circle', 'green', 'red', 'gray', 'yes', 'teal', 'black', 'rectangle', 'yellow', 'triangle', 'brow
n', 'blue', 'no']

--- Reading/processing images...
Read 3000 training images, 1000 validation images and 1000 testing images.
Each image has shape (64, 64, 3).

--- Fitting question tokenizer...
Vocab Size: 27
{'is': 1, 'shape': 2, 'the': 3, 'a': 4, 'image': 5, 'there': 6, 'not': 7, 'what': 8, 'present': 9, 'd
oes': 10, 'contain': 11, 'in': 12, 'color': 13, 'no': 14, 'rectangle': 15, 'circle': 16, 'triangle':
17, 'brown': 18, 'yellow': 19, 'gray': 20, 'teal': 21, 'green': 22, 'black': 23, 'blue': 24, 'red': 2
5, 'of': 26}
```

The details about the dataset is shown in the adjacent picture. The validation images could not be taken at random because of the constraints in the question set. The answer set as well as the vocabulary set is indexed as shown.

The frequency of each possible answer is shown in the adjacent picture. As we can see, majority of the questions are binary. Questions related to shapes comes next followed by colors.



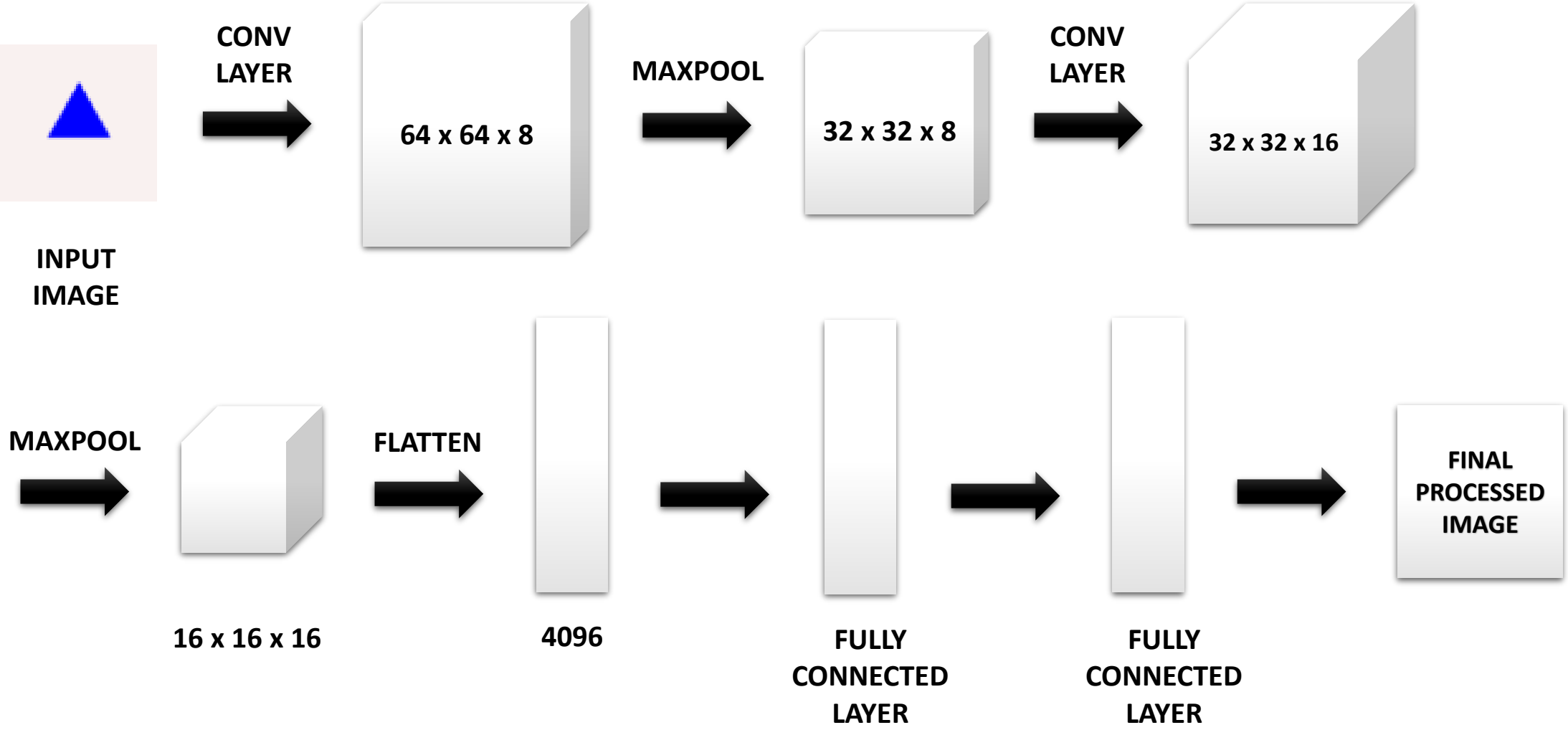
THE APPROACH



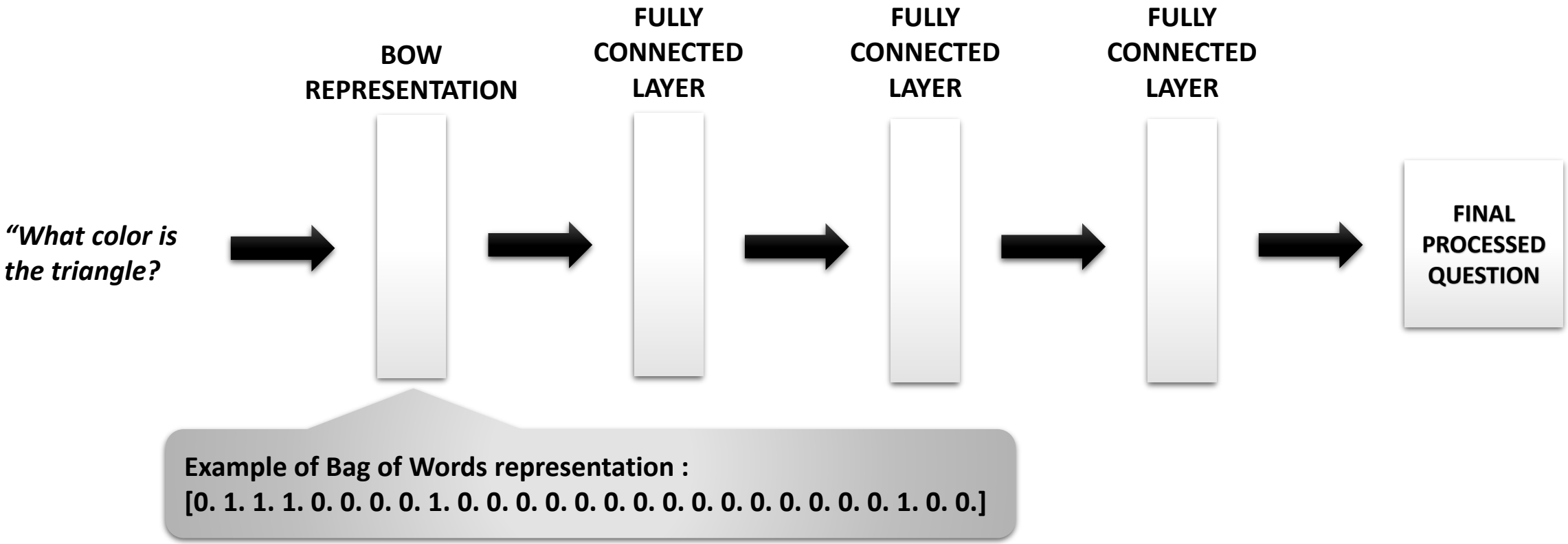
- Processing the image using a **CNN** model
- Processing the image using **Bag of Words** (**BoW**) model
- Combining the above models together to form **the working** model
- Predicting and returning the answer having the **highest probability** (using **softmax)**

The adjacent animation shows an example where the image is of a **blue triangle**, the question being asked is **what is the color of the triangle**. After processing the image and question, the working model predicts the answer as "**blue**".

EASY – VQA ARCHITECTURE

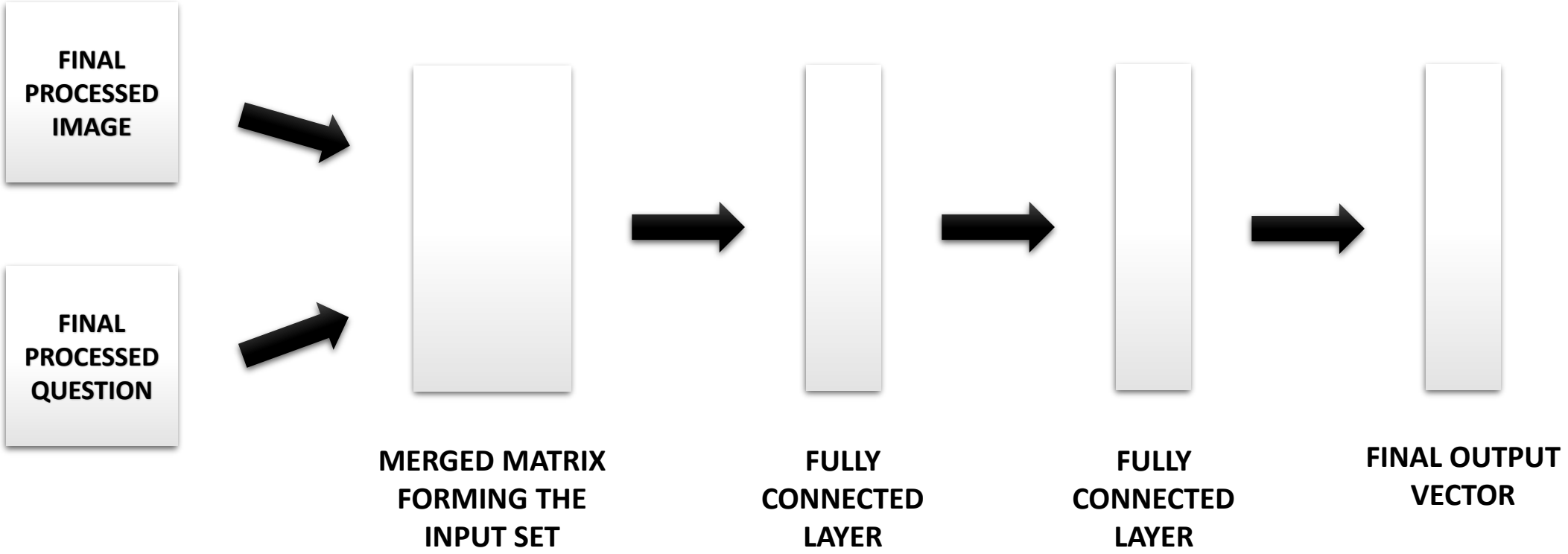


EASY – VQA ARCHITECTURE

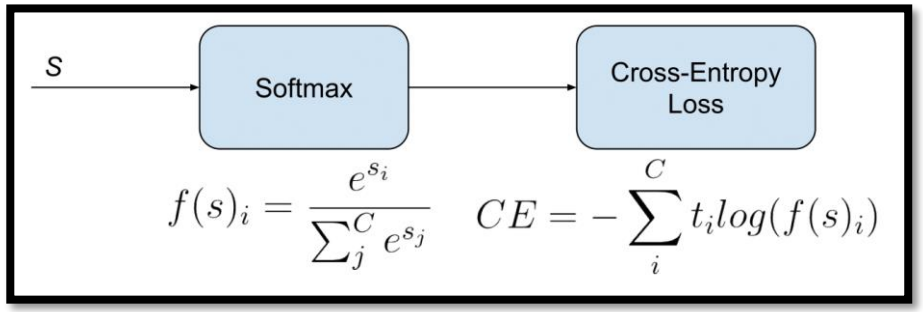


- We have used **Keras** Tokenizer to implement the **Bag of Words** (**BoW**).
- The **array** represents the **words of the entire vocabulary**, each word having a particular **index**.
- A **combination** of these words forms each question.
- **Index positions** that turn to **1** represents the words present in the asked question.

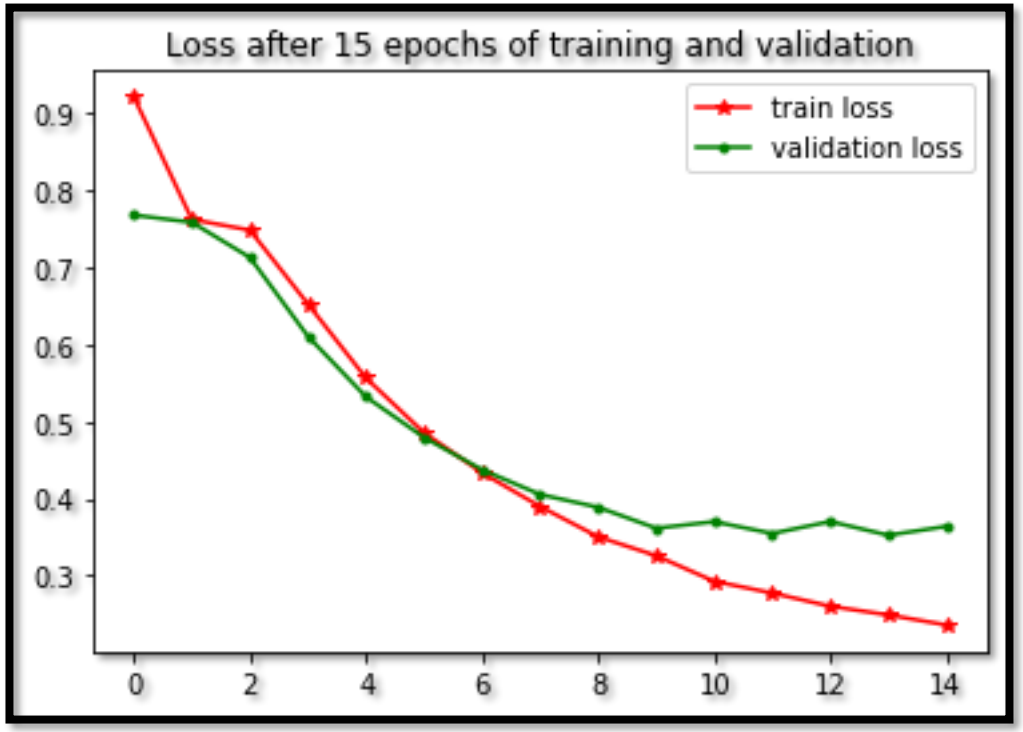
EASY – VQA ARCHITECTURE

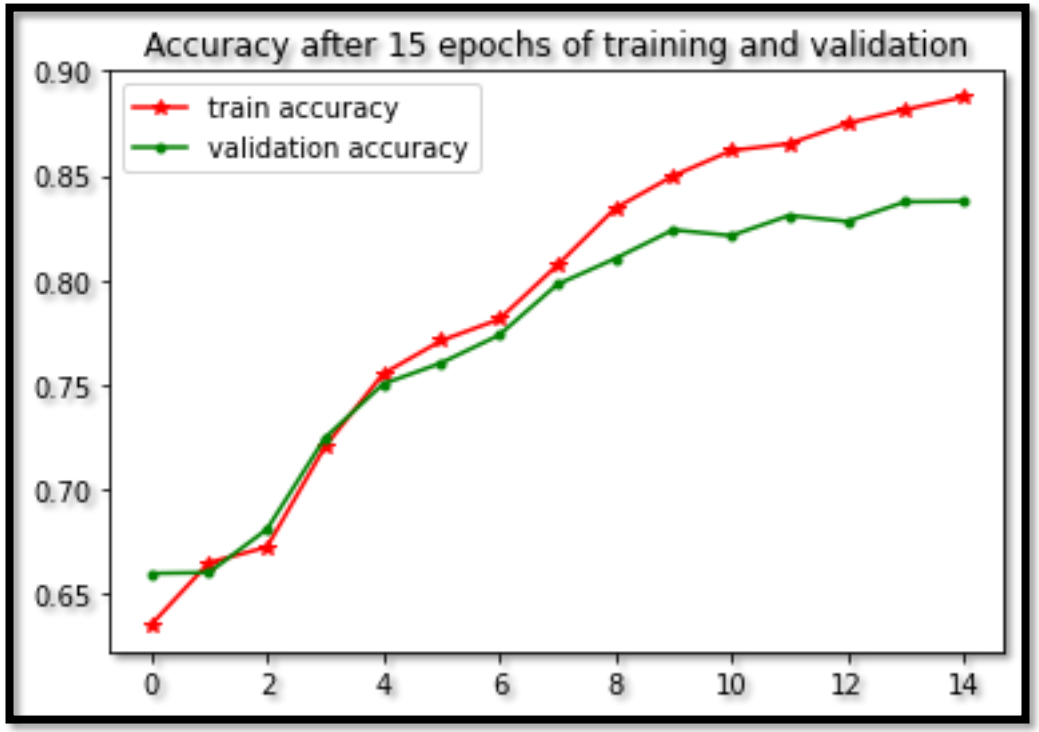


- We have used the **categorical cross-entropy loss** function that is used to multi-class classification.
- The optimizer being used is **Adam** with a learning rate of 5 x 10^{-4}.



EXPERIMENTS AND RESULTS

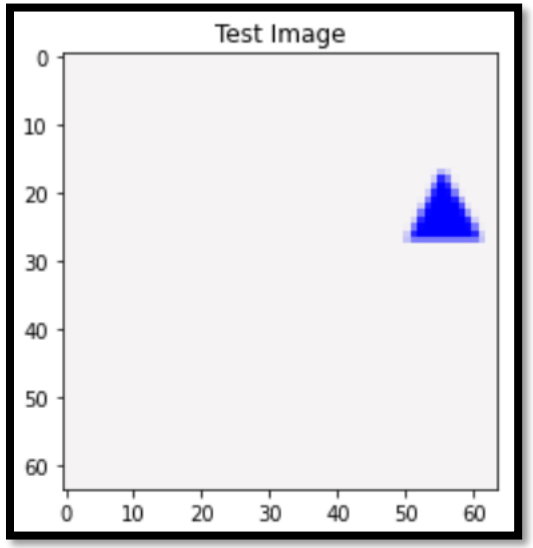




- The model achieved **88.75% accuracy** on the **training set** and **83.76% accuracy** on the **validation set** (when the plots were made)
- The training accuracy remains around **90%** whereas the validation accuracy remains around **84%**.
- After 14 to 15 epochs, the model **overfits**, resulting in increasing the **validation loss**.

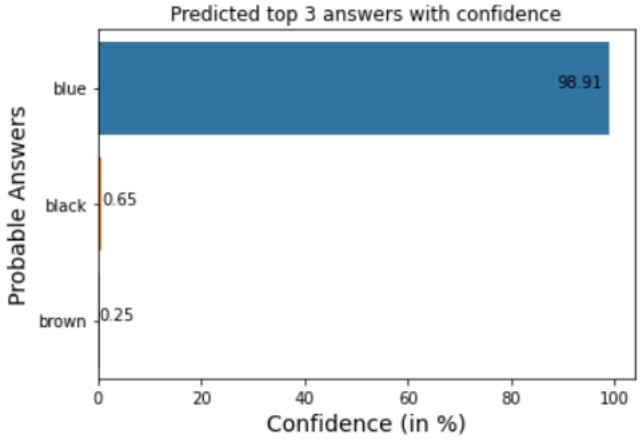
EXPERIMENTS AND RESULTS

INPUT



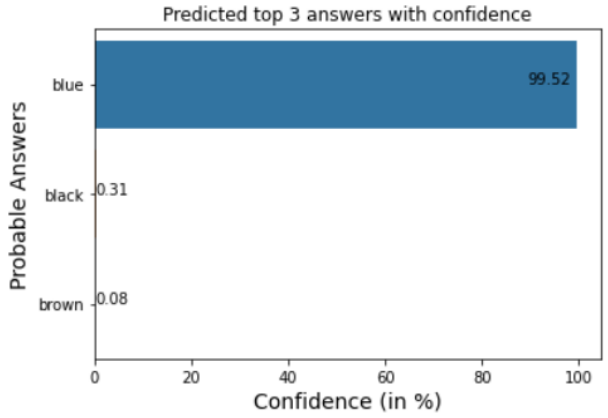
OUTPUT TYPE 1

```
If you want to ask a question from a pre-defined set of questions, enter 1
If you want to ask a question on your own, enter 2
Enter your value here : 1
Pre-defined questions related to this image are as follows :
                                    Questions
0         what shape does the image contain?
1                what color is the triangle?
2         what is the color of the triangle?
3           does the image contain a circle?
4        is there not a circle in the image?
5                      is no circle present?
6                       is there a triangle?
7       is there a black shape in the image?
8     is there not a red shape in the image?
9   is there not a green shape in the image?
10                    is there a blue shape?
11     does the image contain a brown shape?
Enter a question index : 1
AxesSubplot(0.125,0.125;0.775x0.755)
```



OUTPUT TYPE 2

```
If you want to ask a question from a pre-defined set of questions, enter 1
If you want to ask a question on your own, enter 2
Enter your value here : 2
Enter your question here : What is the color of the shape present?
AxesSubplot(0.125,0.125;0.775x0.755)
```



CONCLUSION

To conclude, I would like to discuss about various limitations of my Easy – VQA model.

- The model **could not answer** questions **outside** of its **limited vocabulary of words**. For example, the model identifies the word ‘color’ and not ‘colour’.
- It **does not work** on questions having **multiple outputs**.
- It **gets confused** between the **binary answers** (**yes / no**), returning ‘no’ more frequently for a given question.
- The dataset is way **too small**. Researchers could expand the dataset by increasing the number of images, questions and answers, thereby increasing its limited vocabulary.
- One could implement **RNN** (**Recurrent Neural Network**) to process the input questions.

REFERENCES

- **PROJECT INSPIRATION :** https://visualqa.org/
- **MY PROJECT ADAPTATION FROM :** https://victorzhou.com/blog/easy-vqa/
- **DATASET :** https://github.com/vzhou842/easy-VQA/
- **IMPLEMENTATION IDEA :** https://github.com/vzhou842/easy-VQA-keras/
- **PROJECT DEMO IDEA :** https://easy-vqa-demo.victorzhou.com/

THANK YOU