VISUAL QUESTION-ANSWERING

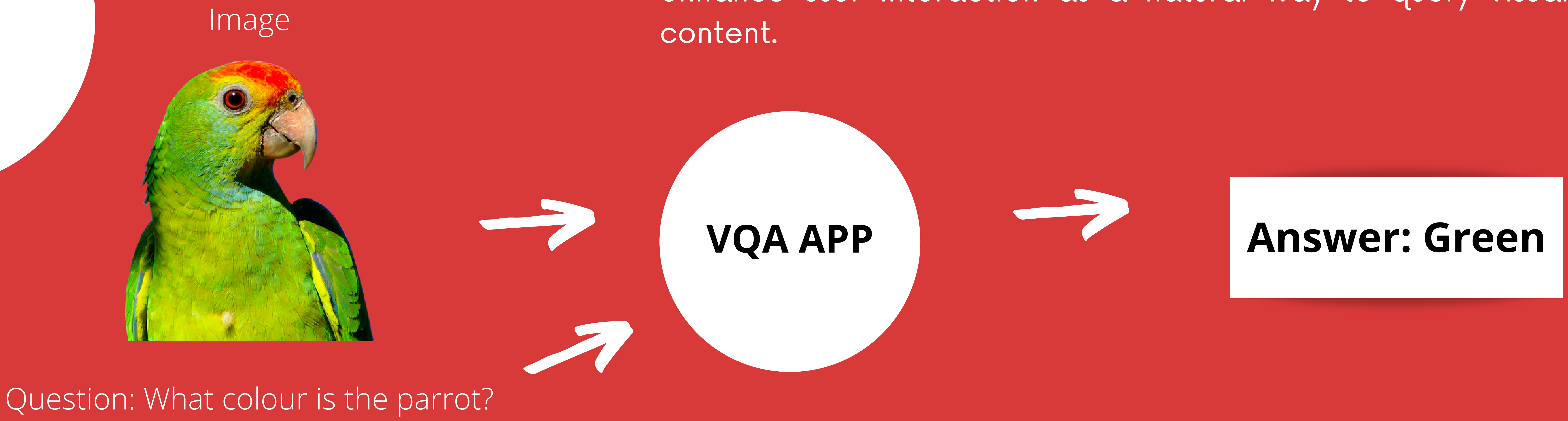
AKSHAY RAJNITH
HARSHITHA KOLUKULURU
JANHAVI KASHYAP
MANISHA R
PADMAJA BHAGWAT

MILESTONE 2

PROBLEM DEFINITION

VQA is a problem at the intersection of Computer vision and NLP that answers text-based questions about images. Natural language questions, given their arbitrary nature, can encompass many sub-problems including but not limited to object detection and recognition, attribute classification and counting.

A robust VQA system capable of answering a wide range of questions can be very useful for a number of applications that enhance user interaction as a natural way to query visual content.



Image

VQA APP

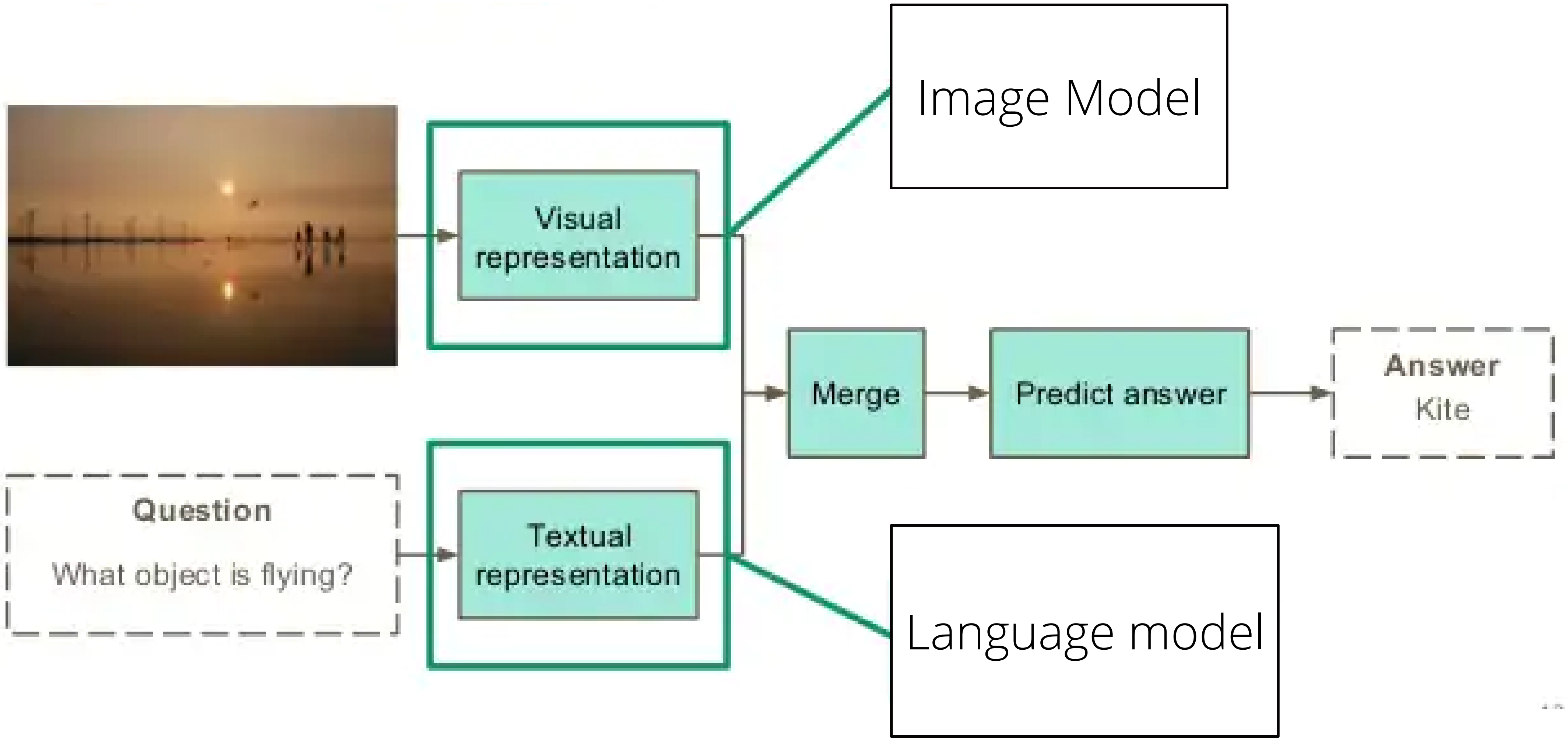
Answer: Green

Question: What colour is the parrot?

PROPOSED SOLUTION

We can build a multimodal model that can take both images and text questions as input and predict the answer.

A high-level flow of the minimum required tasks for this project are given in the diagram



PROJECT SCOPE

I PROOF OF CONCEPT

- EDA
- Baseline models
- Verify how the models do on unseen data
- Visualize model activations to analyse what the model is seeing

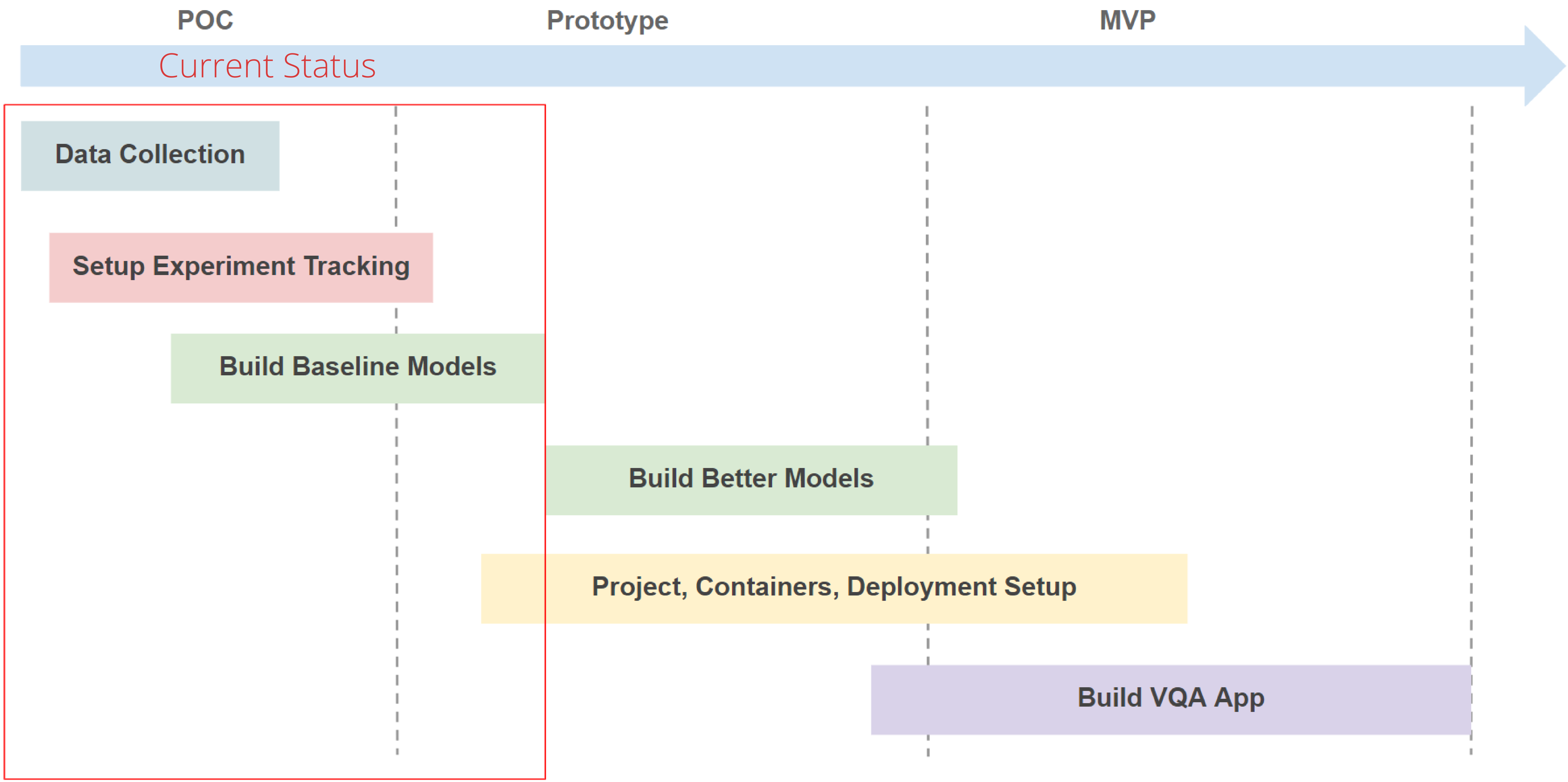
II PROTOTYPE

- Create a mock-up of screens to see what the app would look like
- Deploy one model to Fast API to service model predictions as an API

III MINIMUM VIABLE PRODUCT

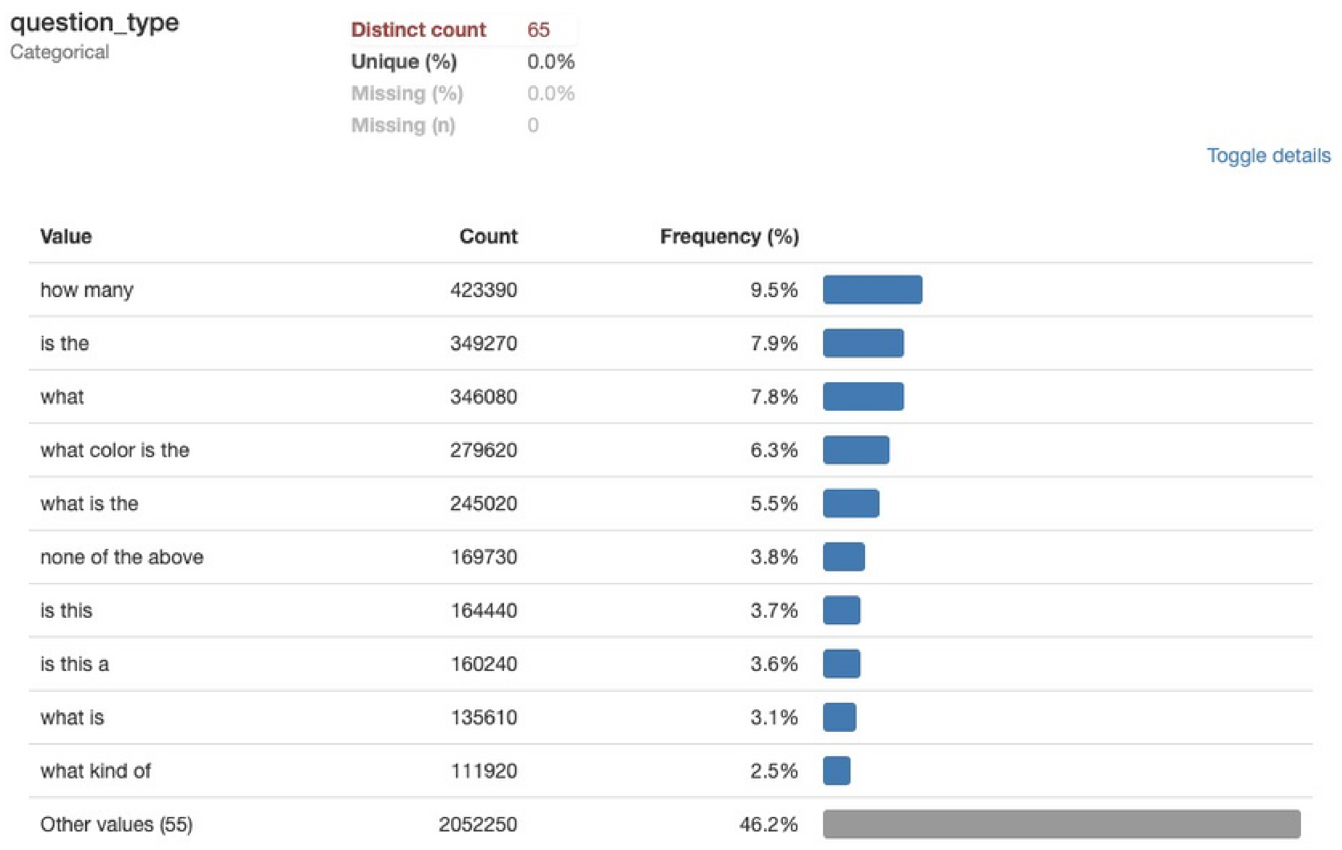
- Create an app that performs VQA
- API Server for uploading images and answering questions

PROJECT WORKFLOW



EDA

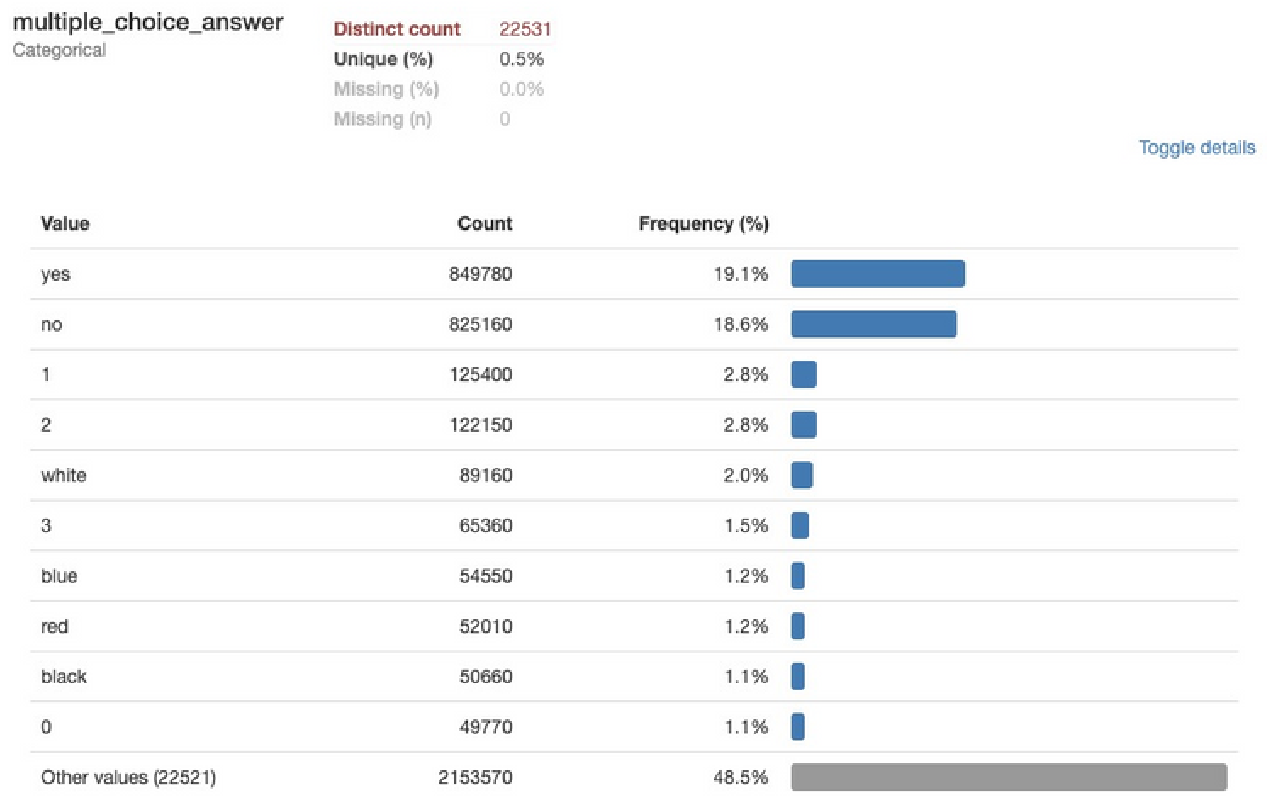
EDA was performed to know the data better



question_type
Categorical

Distinct count	65
Unique (%)	0.0%
Missing (%)	0.0%
Missing (n)	0

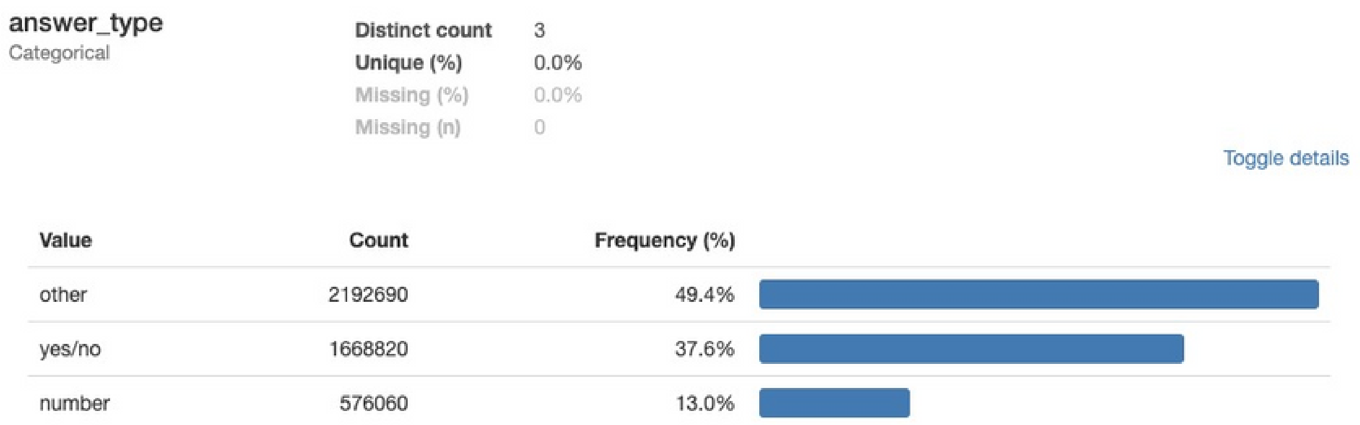
Value	Count	Frequency (%)
how many	423390	9.5%
is the	349270	7.9%
what	346080	7.8%
what color is the	279620	6.3%
what is the	245020	5.5%
none of the above	169730	3.8%
is this	164440	3.7%
is this a	160240	3.6%
what is	135610	3.1%
what kind of	111920	2.5%
Other values (55)	2052250	46.2%



multiple_choice_answer
Categorical

Distinct count	22531
Unique (%)	0.5%
Missing (%)	0.0%
Missing (n)	0

Value	Count	Frequency (%)
yes	849780	19.1%
no	825160	18.6%
1	125400	2.8%
2	122150	2.8%
white	89160	2.0%
3	65360	1.5%
blue	54550	1.2%
red	52010	1.2%
black	50660	1.1%
0	49770	1.1%
Other values (22521)	2153570	48.5%



answer_type
Categorical

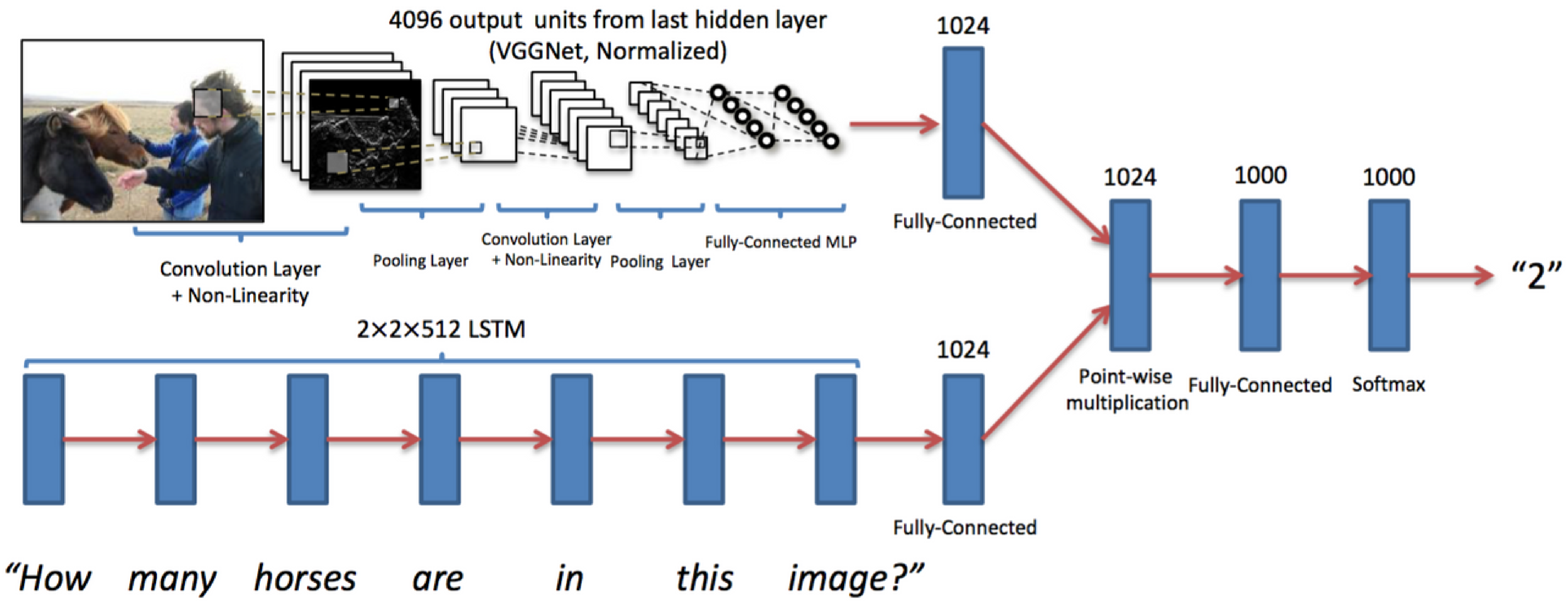
Distinct count	3
Unique (%)	0.0%
Missing (%)	0.0%
Missing (n)	0

Value	Count	Frequency (%)
other	2192690	49.4%
yes/no	1668820	37.6%
number	576060	13.0%

BASELINE MODEL

For our baseline model, we trained a CNN for image feature extraction while the text features were extracted using RNN.

We used a subset of the VQA dataset consisting of 2000 images and corresponding text data.


4096 output units from last hidden layer
(VGGNet, Normalized)
Convolution Layer + Non-Linearity
Pooling Layer
Convolution Layer + Non-Linearity
Pooling Layer
Fully-Connected MLP
1024
Fully-Connected
1024
1000
1000
"2"
Point-wise multiplication
Fully-Connected
Softmax
2×2×512 LSTM
1024
Fully-Connected
"How many horses are in this image?"