Visual Question Answering

Nikhil Gupta

Problem Statement

- Given an image and a natural language question about the image, the task is to provide an accurate natural language answer.

- Visual questions selectively target different areas of an image, including background details and underlying context.

What is the mustache made of?

AI System

bananas

Visualization of Task

VQA

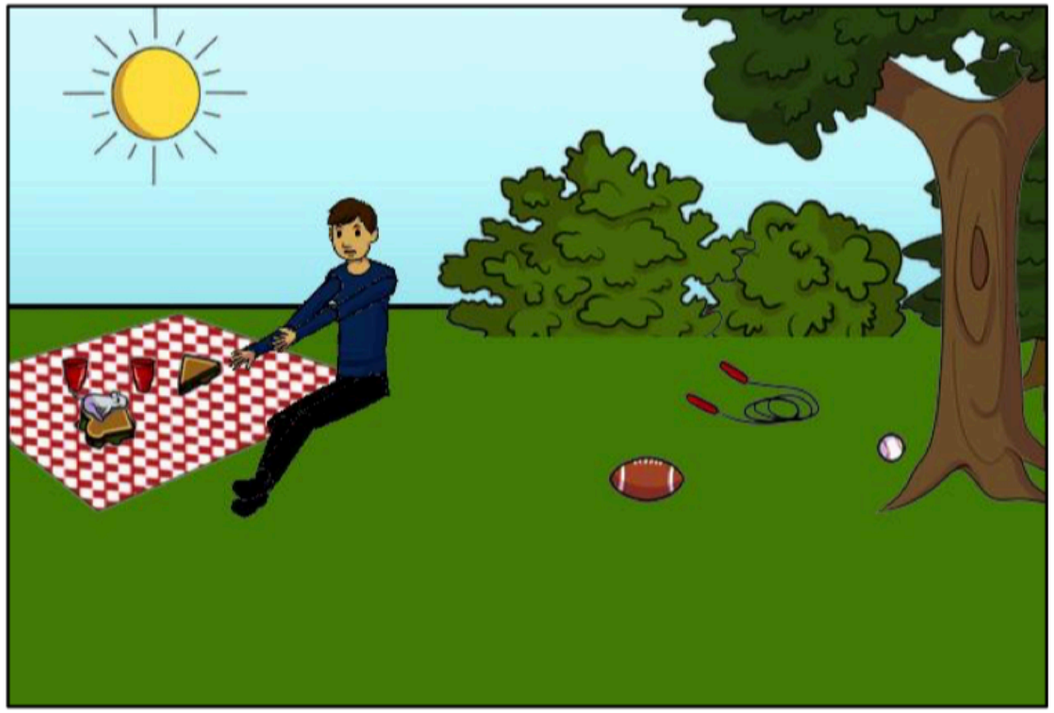
- [ ] Annular Competition held by collaboration of Virginia and Georgia Tech
- [ ] Provide a large data set which acts as a standard to compare accuracies

Data

- The Dataset is split into three sets:
  - Train
  - Validation
  - Test
- Total 1.1 million questions and 25k images
- Each question has been made with context to a specific image

What color are her eyes?
What is the mustache made of?

How many slices of pizza are there?
Is this a vegetarian pizza?



Is this person expecting company?
What is just under the tree?

Does it appear to be rainy?
Does this person have 20/20 vision?

My Work

- [ ] Tried to implement paper for this task:
  - [ ] VQA: Visual Question Answering
    - [ ] Aishwarya Agrawal et Al.
  - [ ] Required a basic knowledge of Machine Learning

The Model

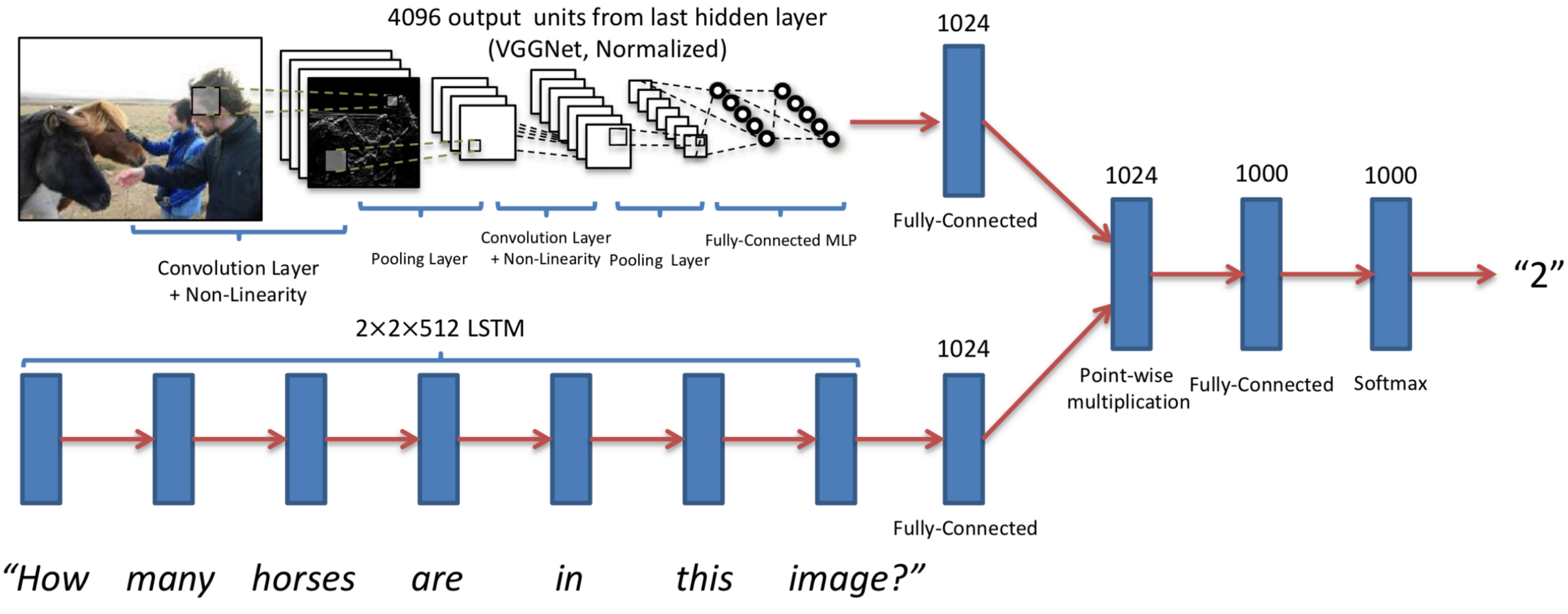
- Consists of two correlated architectures
  - A CNN model (To process the image)
  - A LSTM model (To process the question)
  - Both outputs P.W.M. and sent through fully connected neural networks layers (MLP)
    - 1024 * 1000 and 1000 * 1000
  - Final output obtained from output layer after softmax

CNN Model

- [ ] Used VGGnet (16 convolutional layers)
- [ ] Produced output of 4096 dimensions
- [ ] Passed though fully connected layer (4096 * 1024)

LSTM Model

- Used bi-layer LSTM model with 512 dimension hidden dimensions
- Input words embedded using GLOVE embeddings
- Output of 2048 dimensions
  - 2*512 (2 hidden states)
  - 2*512 (2 cell states)
- Passed though fully connected layer (2048 * 1024)



The Model Visualised

Results

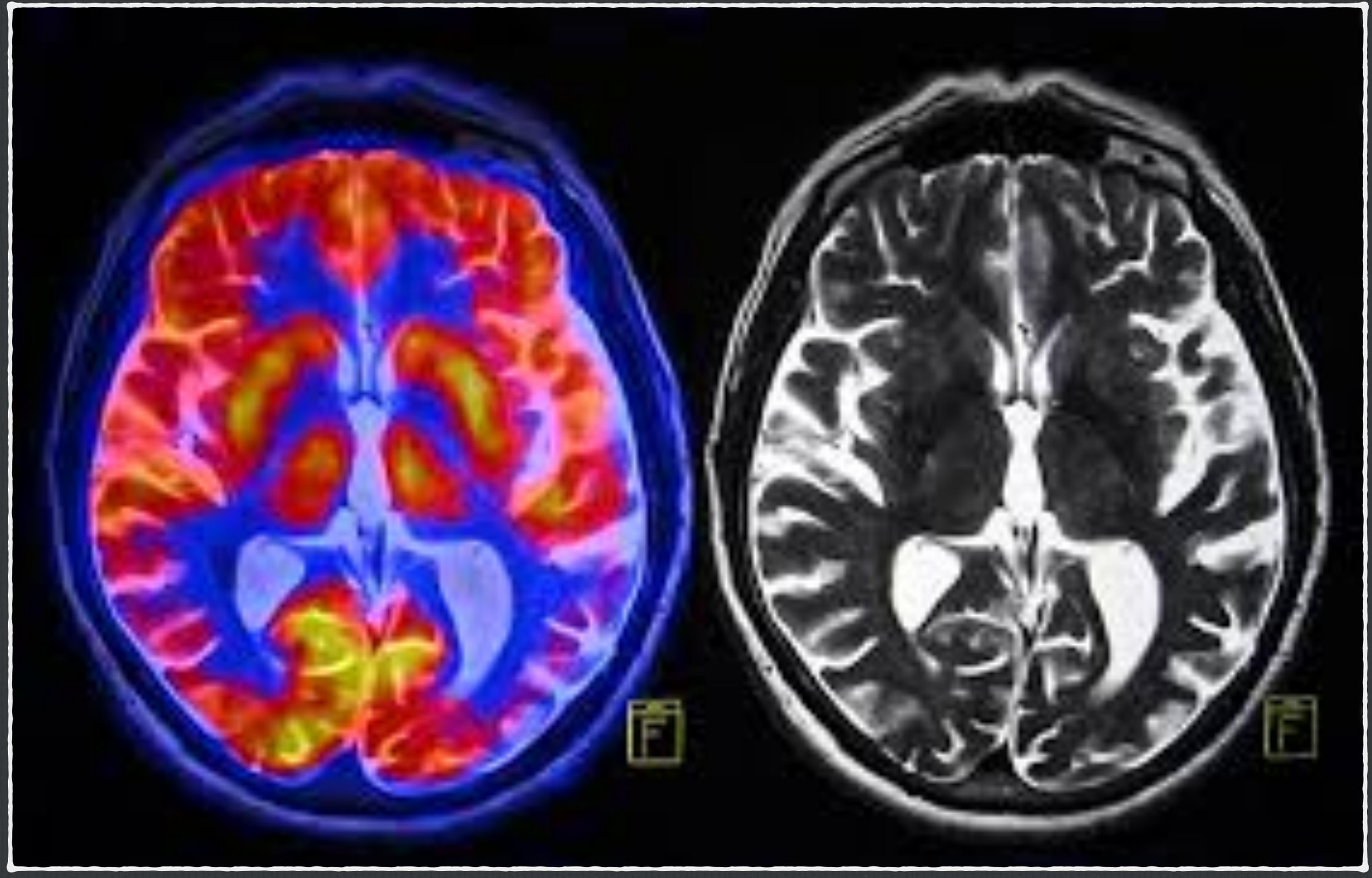
- Able to predict the correct answer with accuracy ~40 %
- Very strong in predicting simpler answers like:
  - “Yes”, “No”
  - “1”, “2”, “3”
  - “Red”, “Blue”, White”

Applications

- There are lots of domains in which Visual Question Answering can be implemented
- Telemedicine
- General Image Analysis
  - Geo-tracking
  - Vandalism Detection

Telemedicine

- [ ] Doctors can send MRI and CT Scan images to doctors off shore
- [ ] Basic VQA will be able to extract useful information from scan and summarise it for doctor
- [ ] Off-Shore doctors can quickly analyse summary and give their diagnosis.



☐ Is there a tumour in the brain?

Thank You