OBJECT AND ACTION RECOGNIZER FOR THE BLIND

Presented by:
Ashwani Kumar(17JE003253)
Shubham Mishra (17JE003230)

Need

- Visually impaired people face troubles due to inaccessible infrastructure and social challenges in daily life. To increase the life quality of those people, we report a portable and user-friendly smartphone-based platform capable of generating captions and text descriptions, including the option of a narrator, using image obtained from a smartphone camera.

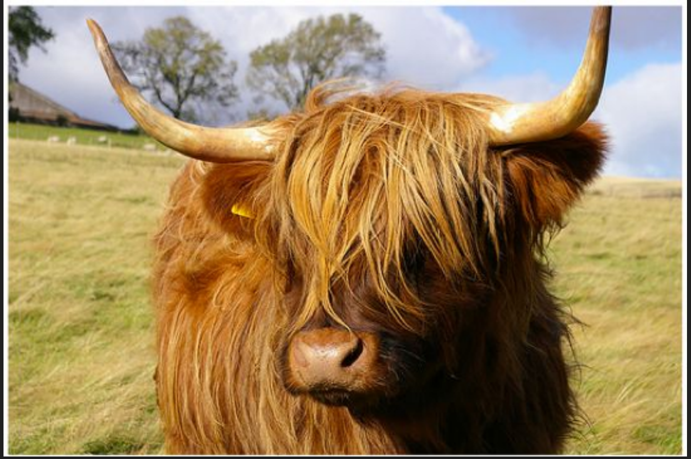
- Our project has great potential to be used for image captioning by visually and hearing impaired people with advantages such as portability, simple operation and rapid response.

Problem Statement

- The image captioning system is integrated with our custom- designed Android application, named as “Let’s Caption It!" which displays and speaks out a short description of the images .

This will help the visually impaired to get an idea about his surroundings.

Adding Captions to Images



- A cow is standing in the field
- This is a close up of a brown cow.
- There is a brown cow with long hair and two horns.
- There are two trees and a cloud in the background of a field with a large cow in it.

Dataset

We used the MS-COCO dataset to train our model.

The dataset contains over 82,000 images, each of which has at least 5 different caption annotations.

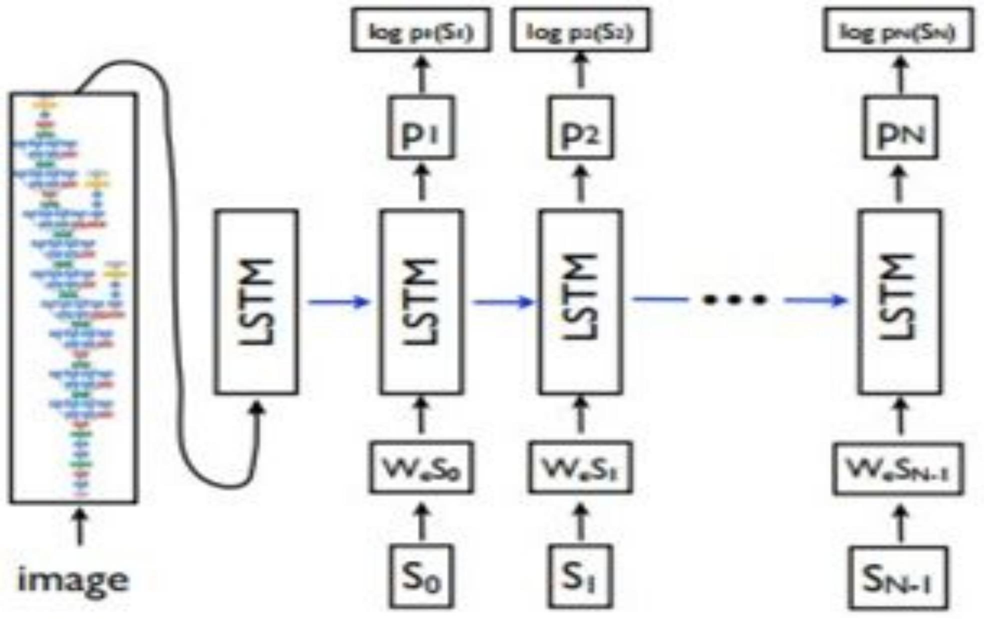
We train our model using these 5 captions, for each image.

Model Overview

- The model proposed takes an image as input and is trained to maximize the probability of sequence of words generated from the model and each word is generated from a dictionary built from the training dataset.

- The input image is fed into a deep vision Convolutional Neural Network (CNN) which helps in detecting the objects present in the image. The image encodings are passed onto the Language Generating Recurrent Neural Network (RNN) which helps in generating a meaningful sentence for the image.

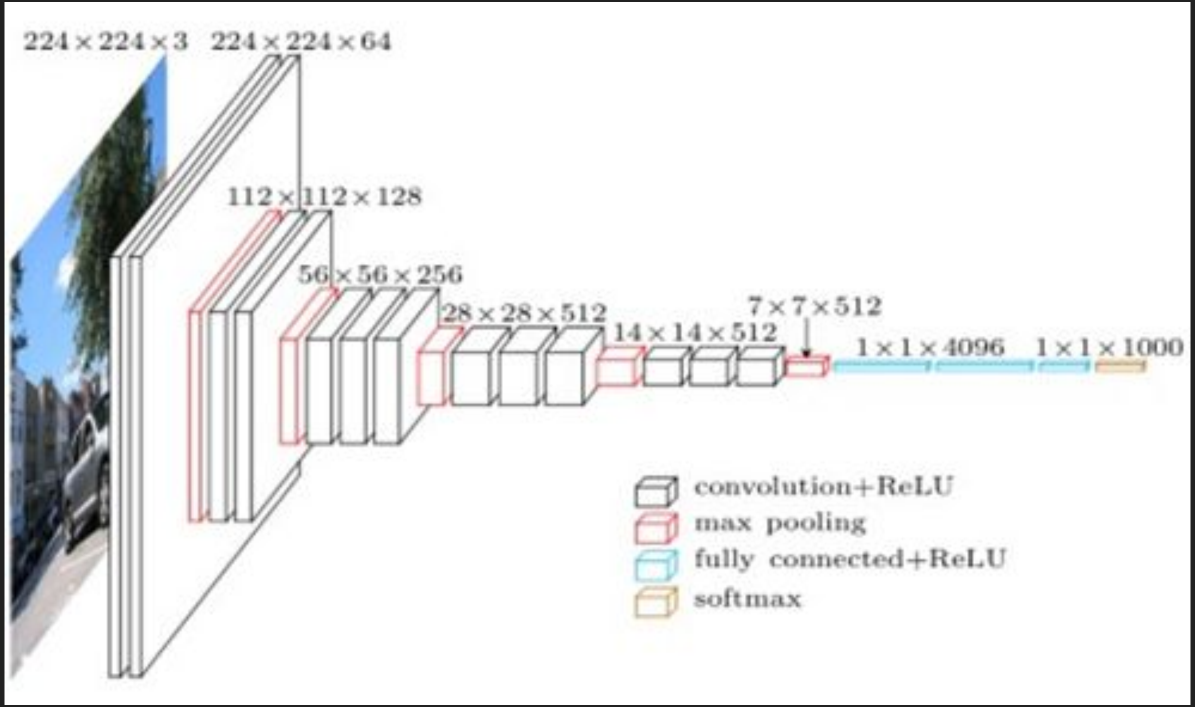
VGG16 + LSTM Decoder



CNN Model

The image taken for classification needs to be a 224*224 image. The only preprocessing done is by subtracting the mean RGB values from each pixel determined from the training images. The image is passed through VGG16 architecture. The output of our CNN encoder is 1*1*4096 encoded which is then passed to the language generating RNN.

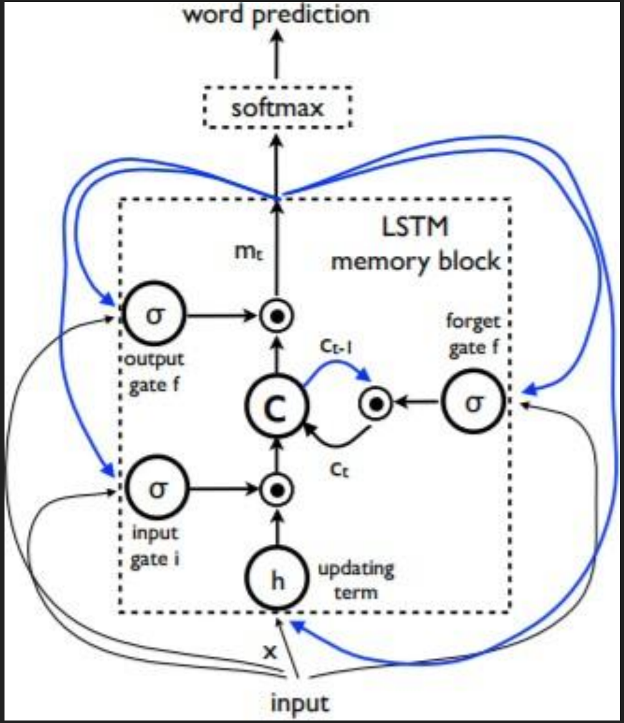
VGG-16 Architecture



LSTM Decoder

- The CNN Encoded vector is passed into LSTM, which uses this to predict the words of sequence. In LSTM decoder, the next hidden state is calculated which is passed through a linear layer and softmax activation to obtain the probability of occurrence of each word.

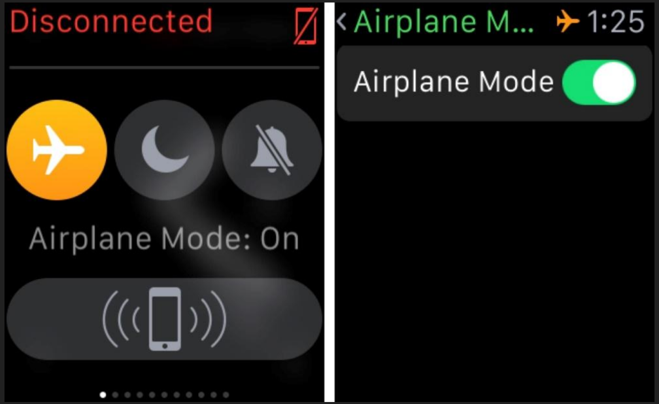
- This process is repeated till end token (.) is encountered by the network. The series of these word prediction generate the caption for the given image.



Working of Android

The blind man can take an image of the surrounding through his camera and using our saved trained models we identify a suitable caption for it. The android mobile then speaks out aloud the caption generated and blind man will be able to hear it and visualize the environment.

An important feature of our app than it can work in airplane/offline mode which is very useful in low internet bandwidth areas and where there is no signal.



Airplane Mode: On

Airplane Mode

Internal Working of Android App

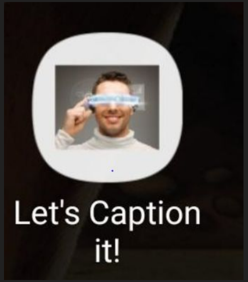
The main activity of the android app has 2 buttons. The button on the right is used to click a image in front of you using a phone camera.

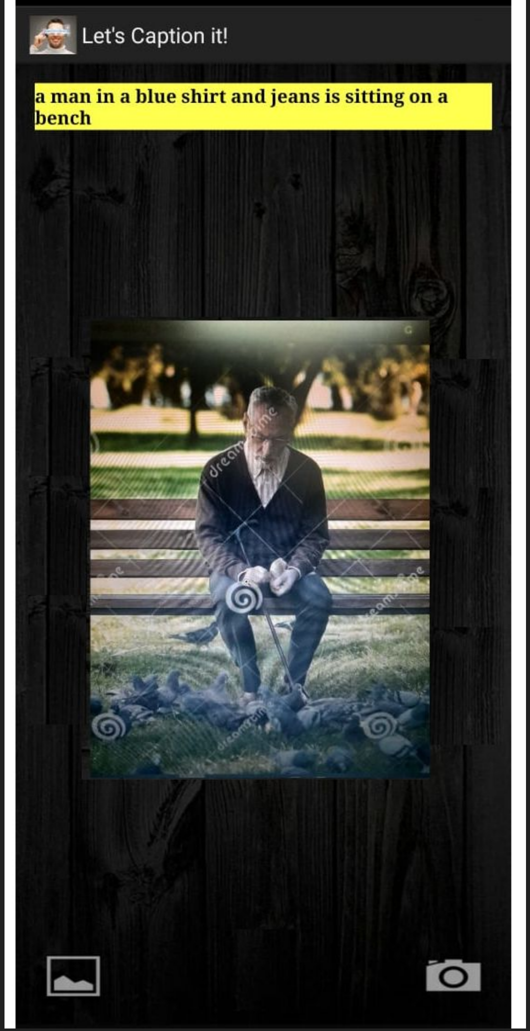
On clicking on the right button the camera intent button and you can capture the image. Additional functionality such as zoom is also available which capturing image.

On clicking on the left button you get an option to select an existing Image.

After you select/click the image, the automatically generates a caption for you an displays in the textview situated in the top of the screen.

The tts feature of Android device is then used to speak out loud the description of the image.





a man in a blue shirt and jeans is sitting on a bench

Thank you.