IMAGE CAPTIONING

Akkem Hema Bhargavi
D Jagan Mohan Reddy
Sowmiya TN
Ganesh G
Mentors: Mr. Naveen Nanda & Ms. Akshita Mehta

Abstract

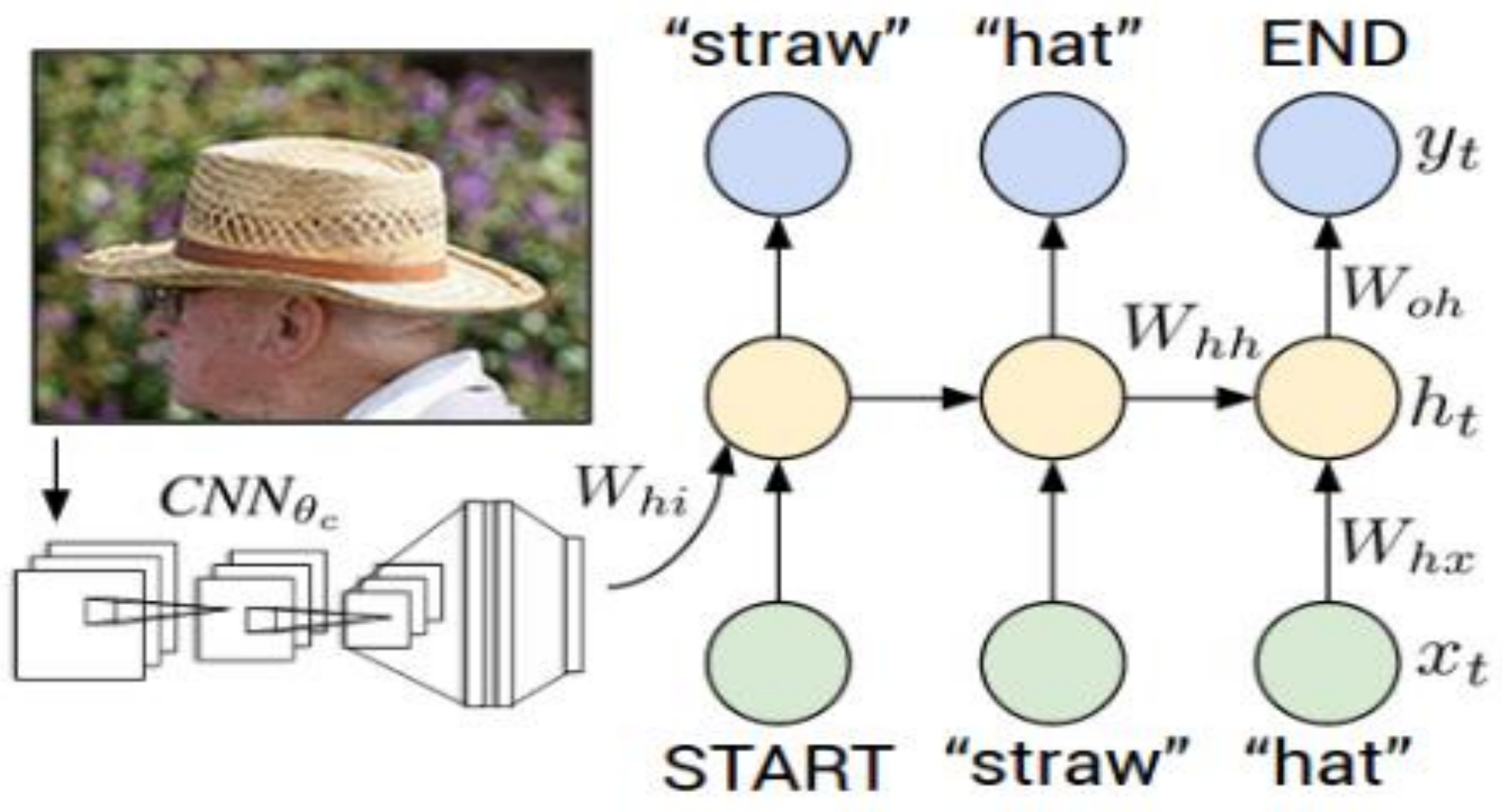
•The purpose of this model is to generate captions for an image.Image captioning aims at generating captions of an image automatically using deep learning techniques.

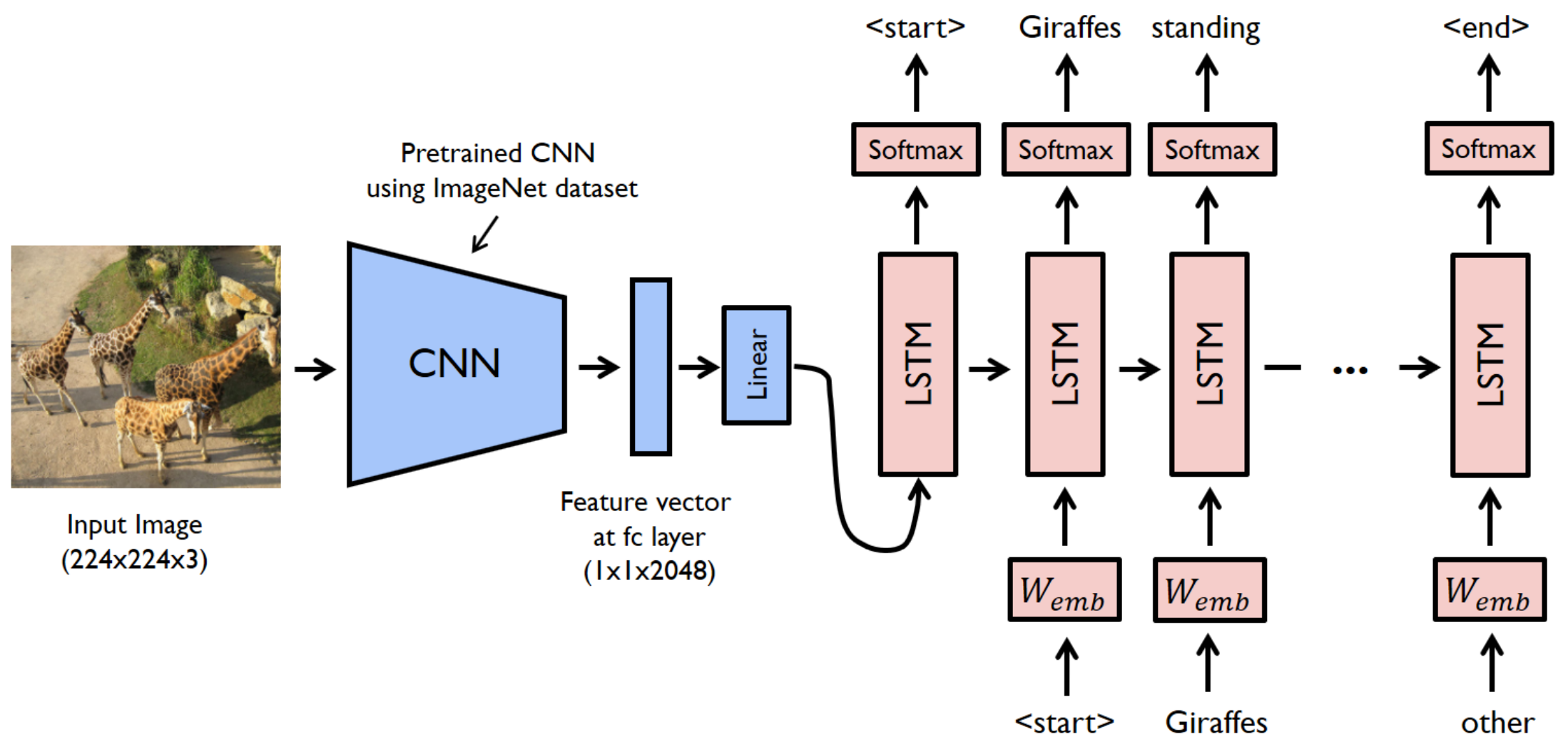
• Initially, the objects in the image are detected using a **Convolutional Neural Network (InceptionV3).** Using the objects detected , a syntactically and semantically correct caption for the image is generated using **Recurrent Neural Networks (LSTM**) with attention mechanism.

• In our project, we are using a traffic sign dataset that is captioned by the above mentioned process .This model is of great benefit to the visually impaired in order to cross roads safely.

Introduction

•**Image captioning** or generating the description of an image in natural language has received a lot of attention in recent times. It emerged as an important and challenging area with research advancing in image recognition. It is interesting due to the fact that it has a lot of practical applications like labelling large image datasets, assisting the visually impaired. It requires the level of understanding way beyond the general object detection and image classification and so, is regarded as a grand challenge. The field is a crossover of the modern Artificial Intelligence models of Natural Language Processing and Computer Vision.





Proposed Method

•We have trained a visual attention based model for the captioning of images with traffic signs. This model can also be extended for a general captioning which might require a larger dataset of images. Our model first pre-processes the images through InceptionV3 pre-trained model. Then, the last output layer is further taken through some convolution layers which is then processed to output the features with dimensions as required.

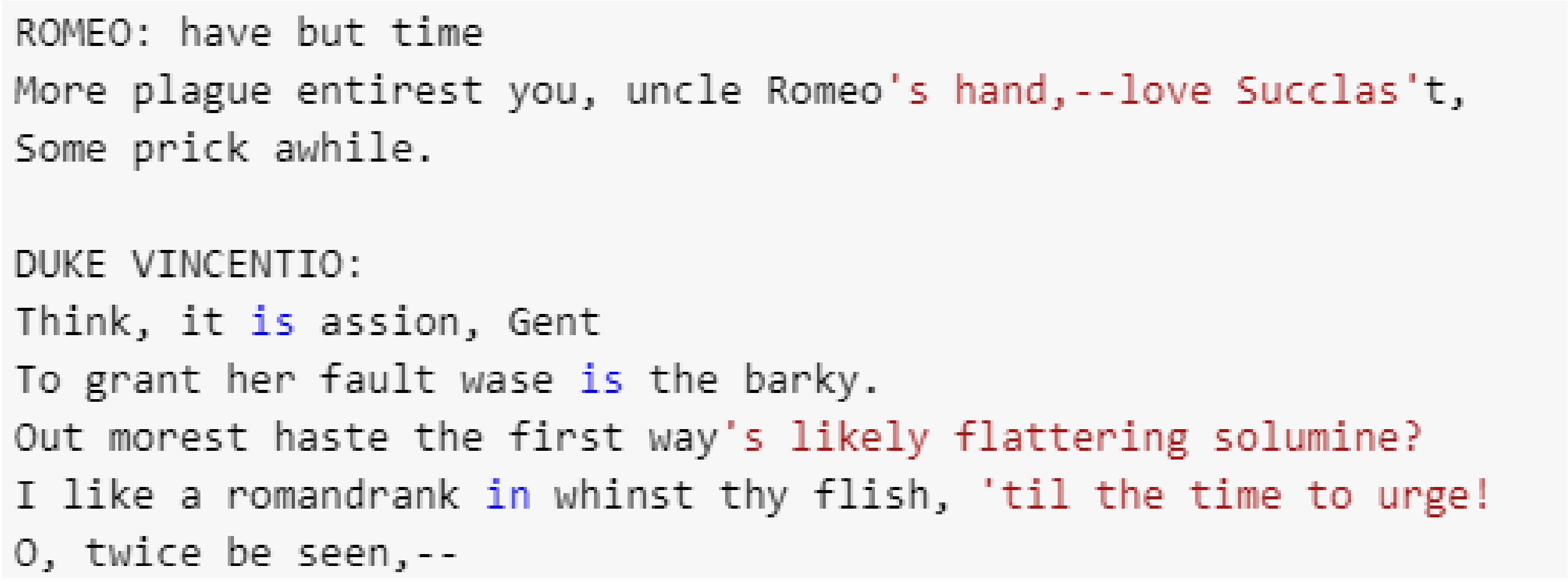
•Parallelly, the captions of every image are tokenized and <start>, <end> tokens are appended for all the sentences. Finally, the model is trained with each image and it's corresponding real captions. The final weights are used to predict the captions for the test images. These captions are not generated at a time but rather word by word. The prediction of every word requires the probability of occurrence of that particular word, given the previous words.

•This method turned out to be quite effective and this is proposed by us from our research.

Experimental Results and Discussion

We have initially run two individual CNN and RNN models. The CNN model was to extract the features of an image and to detect different features in a given image. This was achieved using Mask RCNN and YOLOv3. The results were good even with limited dataset and computational power. Following are the results for both the CNN models.

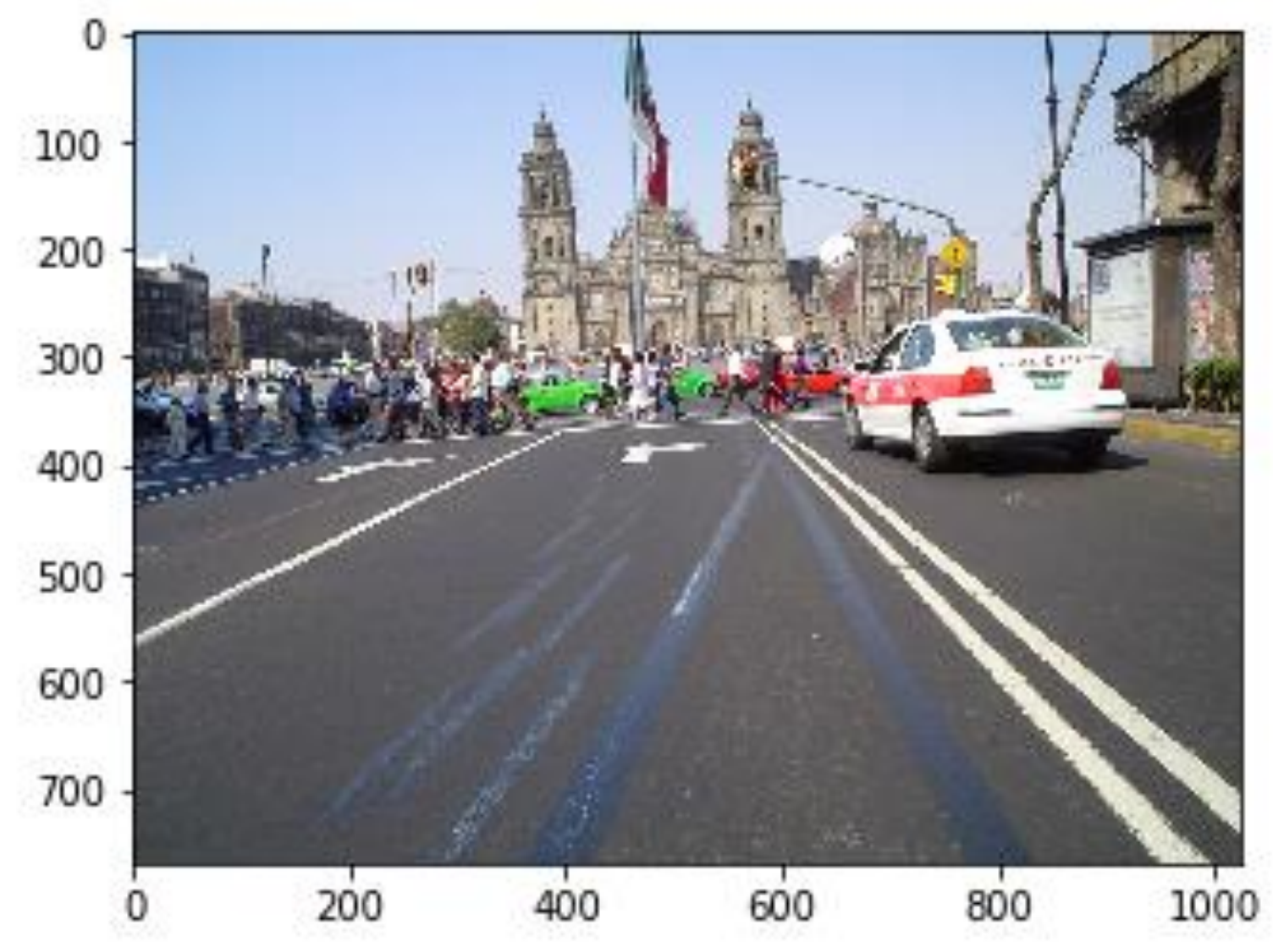
Similar to the CNN models, we have tried out an RNN model which was capable of generating new text from the Shakespeare dataset.



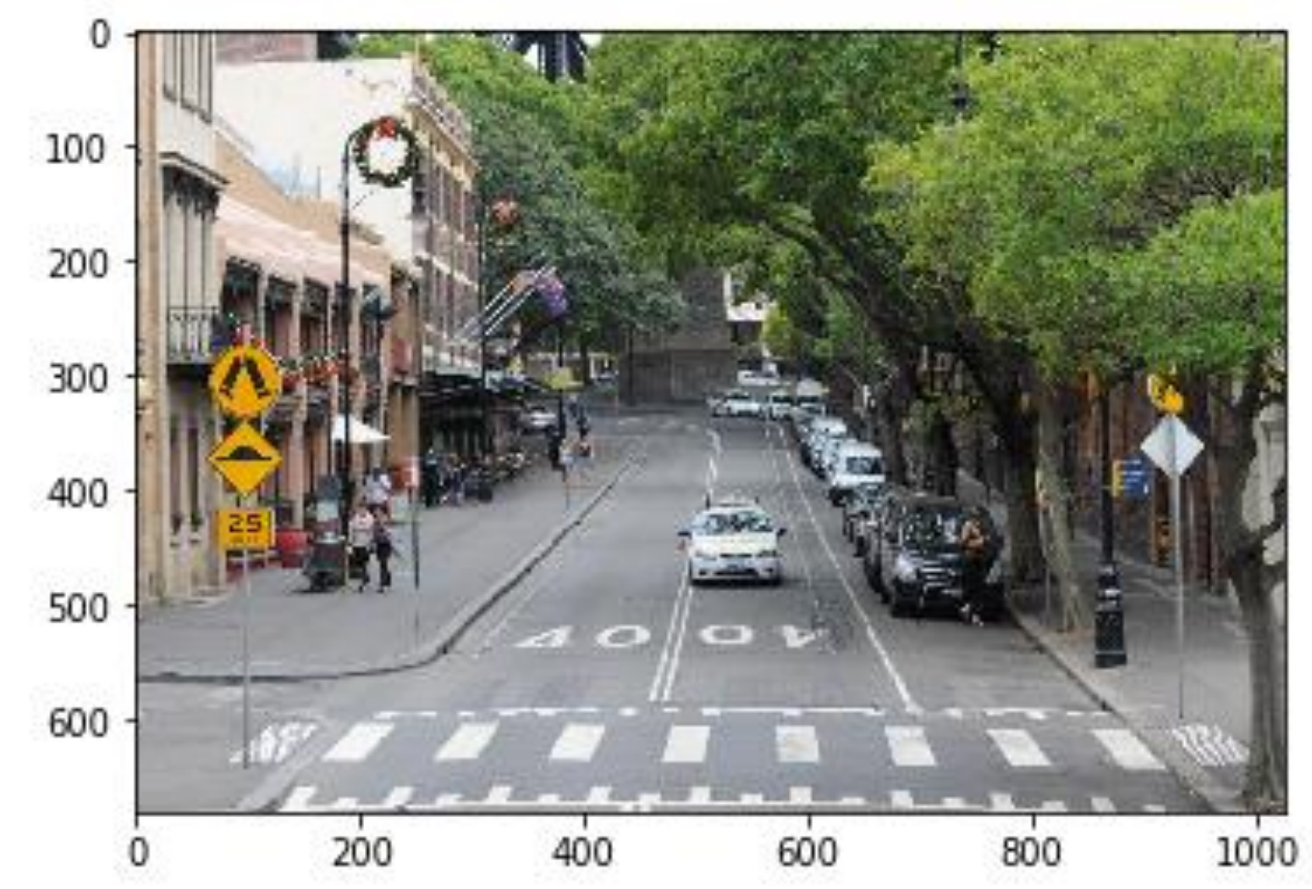
```
ROMEO: have but time
More plague entirest you, uncle Romeo's hand,--love Succlas't,
Some prick awhile.

DUKE VINCENTIO:
Think, it is assion, Gent
To grant her fault wase is the barky.
Out morest haste the first way's likely flattering solumine?
I like a romandrank in whinst thy flish, 'til the time to urge!
O, twice be seen,--
```

Finally, we were able to integrate both the models. We have used a custom made dataset which comprised of the Flickr 8k dataset and some of our custom annotated traffic images. These images were first pre-processed through InceptionV3 architecture which were then passed through some of the convolutional layers to match our required output.

Parallelly, the captions were tokenized and each caption was appended by start and end tokens for the processing of individual captions. The captions were run through the RNN model to generate captions corresponding to the features extracted from a given image using the previous CNN model.



```
Greedy: two people are walking past bus
```



```
Greedy: speed ahead
```

Conclusions

In our project we have developed a model to **caption the images** and convert the **text into speech .**We have done research in order to understand our models in depth and have executed each model separately. We learned how the deep learning techniques work and how to create these models . We faced many challenges while running the model and with our datasets . But later we learnt how to rectify the mistakes and make an efficient model.We also extendded our work by converting the text into voice.
We will be doing future works on voice synthesis.

References

I. Andrej karparthy,Li Fei-Fei**, "**Deep Visuals-Semantic Alignments for generating Image Description", Department of Computer science , Stanford University , Published in IEEE Explore 2015.

II. Oriol Vinyals , Alexander To shey , samy Bengio , Dumitru Erhan ," Show and Tell : A neural Image Caption generator", a rXiy 2015.

III. Rahul Singha , Aayush Sharma 'Image Captioning using Deep Neural Networks" ,Iowa State University , Published in Research gate 2018.

IV. KelvinXu ,JimmyLeiBa ,RyanKiros ,KyunghyunCho ,AaronCourville , RuslanSalakhutdinov, RichardS.Zemel Z,YoshuaBengio, "Show,AttendandTell:NeuralImageCaption GenerationwithVisualAttention",arXiv 2015.

Team No. - 16