# Mini-Project Report


**Team ID: 3**

**Team Members:**

Rachna S Kedigehalli (IMT2019069)

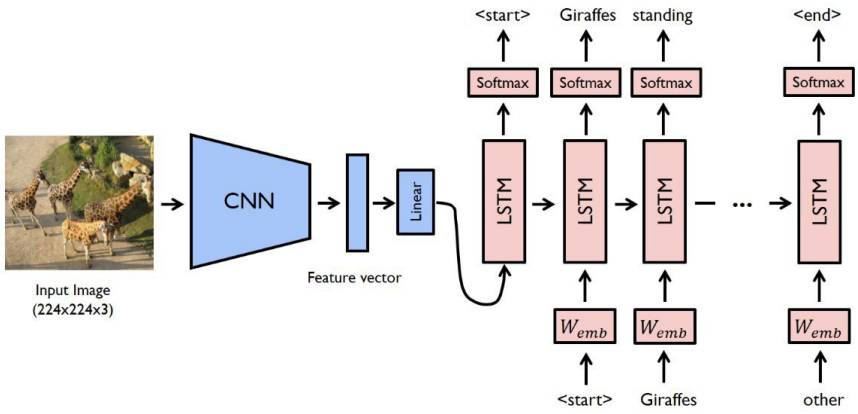
K Yashovardhan Reddy (IMT2019097)

Nandakishore S Menon (IMT2019057)


# Image Captioning

## About Dataset

Flickr8 consists of 8000 unique images chosen from six different Flickr groups, not containing any well-known people or locations. The images were manually selected to depict a variety of scenes and situations. Each image is mapped to up to five different sentences which describe the image. It is optimal for image captioning tasks as it is small in size and thus, the model can be trained easily on low-end devices.

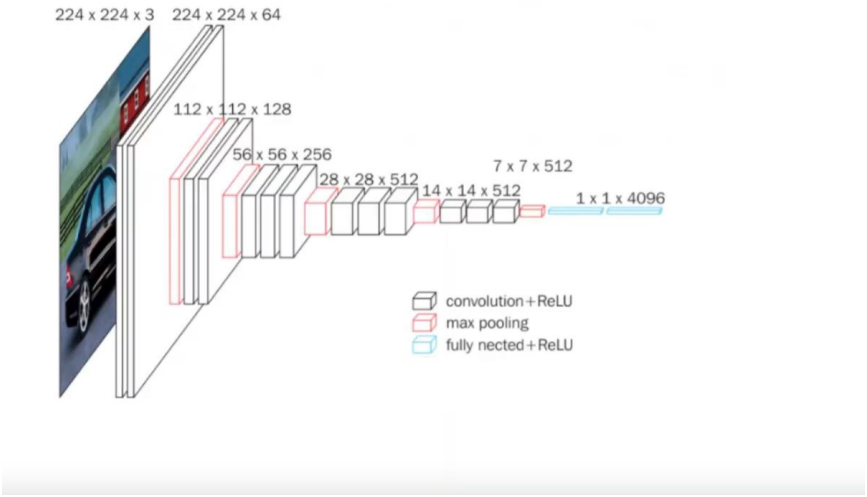



## Architecture Details

- Since the model needs input in the form of vectors and numbers, vocabulary is built and the captions are encoded with a start and end tag.
- Images are transformed to images of shape 224x224.
- In one of our experiments, the images were converted to 1x1x4096 and the word embedding was fixed at a length of 256. The hidden layer size in the LSTM was

256.

- We have experimented with both resnet and inception v3




## Result

BLEU Score with Resnet: `0.05274992063641548`

Average BLEU Score (Inception V3): `0.4582368573517308`

# Language Bias

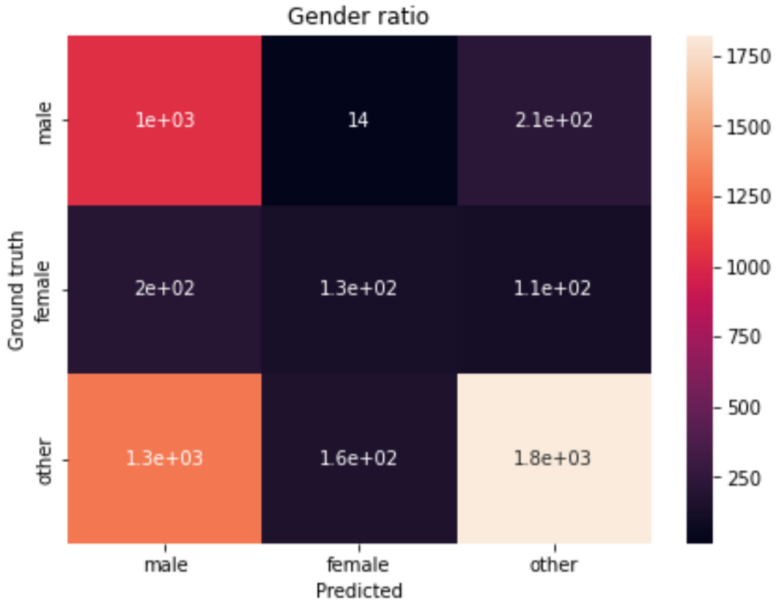
Language bias of an image captioning system is when the system learns from the structure of the language and predicts based on that, without even learning much insight into the context of the image.

## Experiments performed

### Gender Error Rate

We analyze the error rates when describing men and women by looking at the captions generated. A caption is said to be describing men if it has at least one of the male words ( `["male", "man", "boy", "gentleman", "guy"]` ) and none of the female words ( `["female", "woman", "girl", "lady"]` ). Similarly, for female. All other sentences are set to `“other”`.

Below is a heatmap describing the same:




It is seen that the model predicts captions with male words for 1300 instances where the ground truth captions do not have male words in them. This is a very significant number. Similarly, there are 200 instances where the ground truth captions have female words (no male words), but the predicted captions have no female words in them (only male words).

## Gender Ratio

Gender ratio: ratio of captions that include only female words to captions that include only male words. Ideally, these ratios for ground truth captions and predicted captions should match. But we find that the ground truth ratio is `0.3499210110584518`, while the predicted ratio is `0.11984282907662082`.

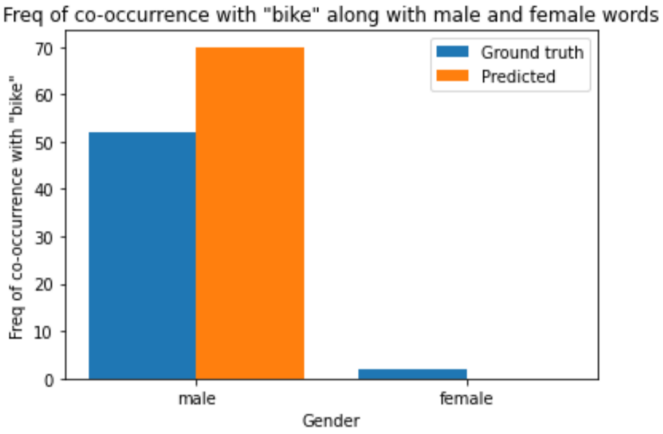
```
Gender ratio of female to male:
Ground truth captions-> 0.3499210110584518
Predicted captions->    0.11984282907662082
```

## Object-Gender Co-Occurrence

We analyze how gender prediction influences prediction of other words or vice-versa. To do this, we plot a bar graph indicating frequency of captions where a word co-occurs with each of the gender words. Ideally, the height of bars for each gender should be equal.

Below is such a plot for the word “bike”. It can be seen that “bike” appears in much higher number of predicted captions along with male words than in ground truth

captions. However, though few ground truth captions have the co-occurrence of “bike” and female words, there are no such predicted captions.




**Conclusion:** From all these experiments, it can be inferred that the model is learning more from the structure of the language than the context of the image.