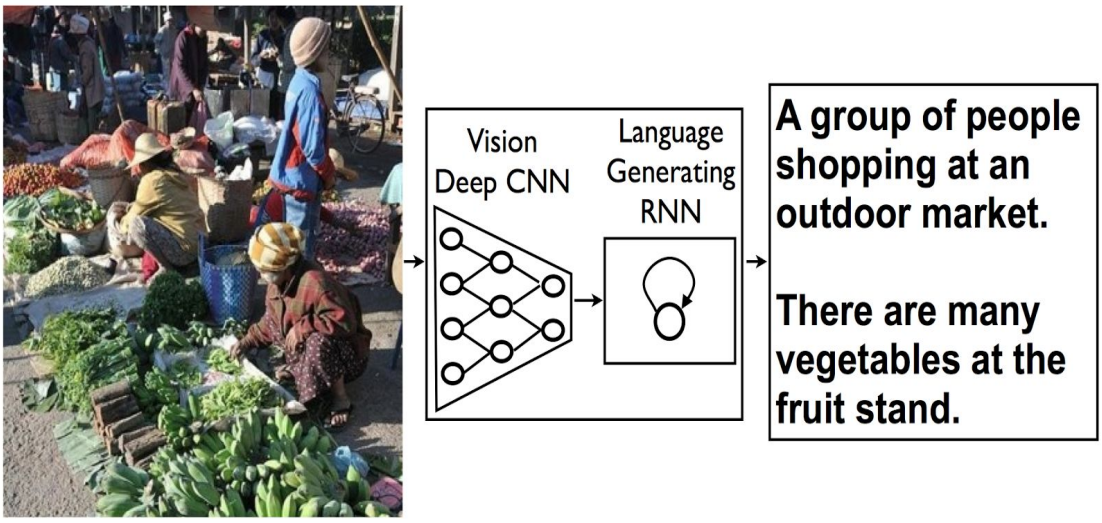
Generate Caption from Image
Project Report



Introduction

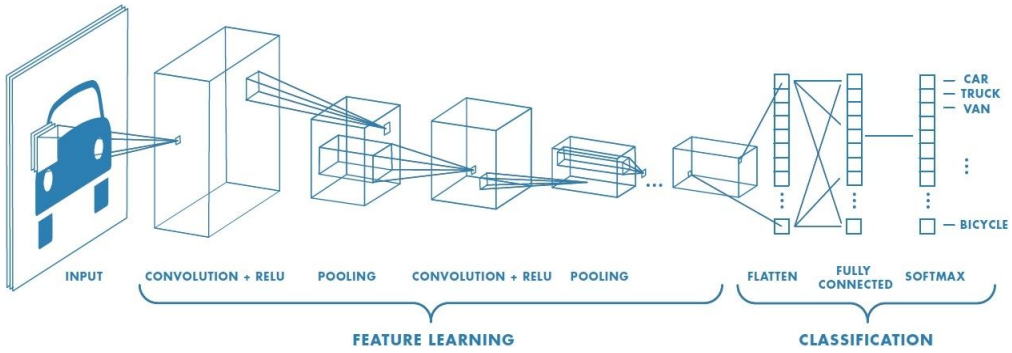
It is easy for people to summarize an image to a few words so that another person can get an idea about the same. This easiness comes from the understanding of things present in the picture. But how will the computer do it without that understanding?

We are using the Convolutional neural network (CNN) and Recurrent neural network (RNN) to get this task done by computers. We used Flicker 8k Data to train our Deep Learning Model. Section 1 contains a brief description of CNN, Inception Model, RNN and LSTM. Next section possesses details of our architecture. The final section includes results and conclusions.

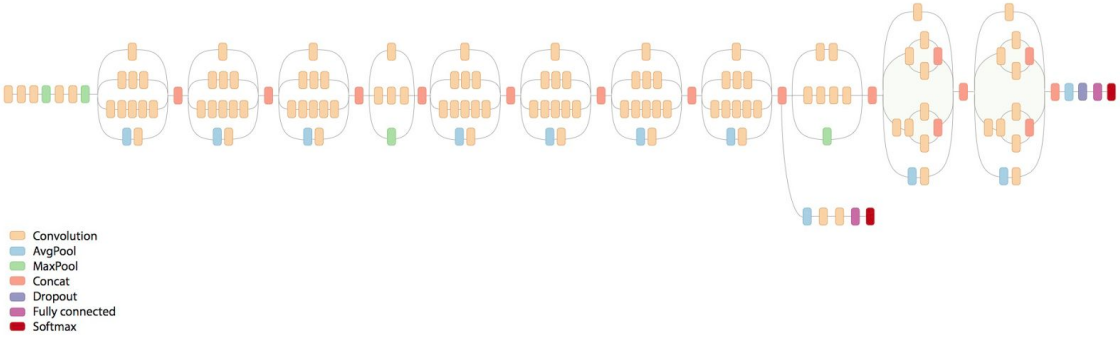
1. CNN, Inception Model, RNN and LSTM

1.1. CNN (Convolutional Neural Networks) - CNN is a kind of artificial neural network (ANN) used to analyze images. Images are high-dimensional vectors. It would take a large number of parameters to create an ANN. CNN solves

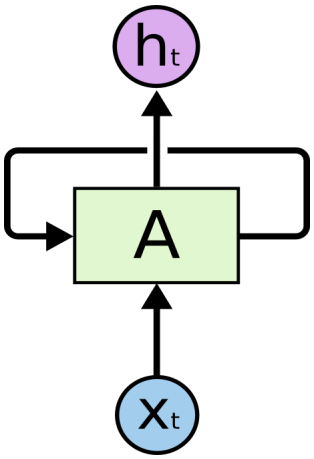
this problem by reducing the number of parameters and adapting the network architecture specifically to vision tasks. CNN is usually composed of a set of layers viz. convolution layer, activation layer, pooling layer, and fully connected layer.



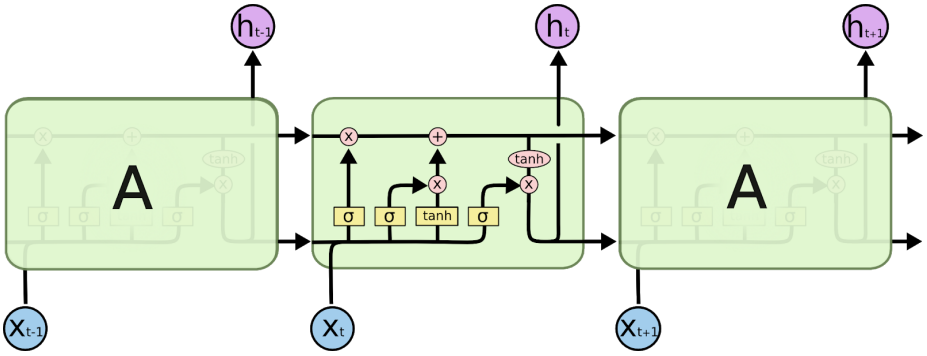
1.2. Inception Model - It is a type of CNN. In this model, many convolution layers exist in parallel and output from all these layers are concatenated to feed into the next layer. Inception v3 is an inception model trained on ImageNet dataset.



1.3. RNN (Recurrent Neural Network) - It is a type of NN which have a loop to make information persistence possible. Data persistence is very useful when our data is sequential because current predictions are dependent on previous ones. RNN allows feeding of past results to the model. So that, our model can extract information from these also to give better predictions.



1.4. LSTM (Long Short-Term Memory networks) - It is a kind of RNN, capable of learning long-term dependencies. It is now widely used in creating models based on a sequential dataset.

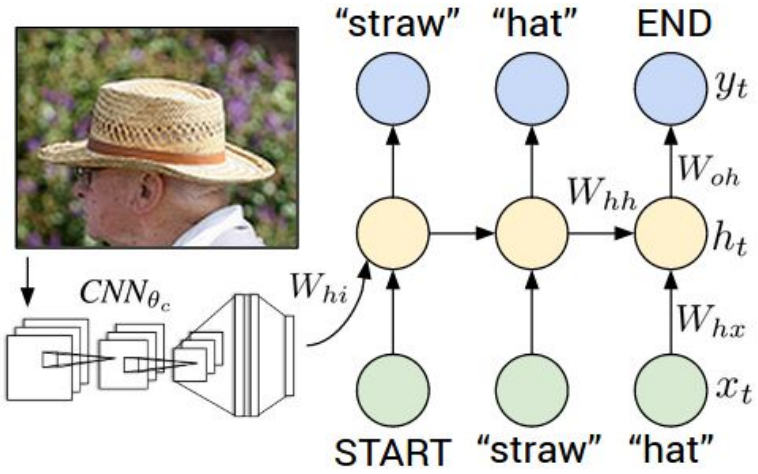


2. Deep Learning Model Architecture

The model has the following two parts:-

1. Encoder - We have used Inception v3 to encode an image to a vector of length 2048. Since Inception v3 model was trained to classify images of different categories, this vector will contain very differentiable features. The dimension of the image is also reduced to have a feasible training time and memory space.

2. Decoder - We have used LSTM to decode 2048 dimensional vector to generate caption of the picture. LSTM allows the use of previous words to produce the next ones. If a part of the generated caption is ***A man is standing in black***, then most probably the next word will be ***shirt***. This feature allows LSTM to generate better captions comparative to simple ANN where previous information is not used.



Results and Conclusion

We have trained our model for 15 epochs which took around 4 hours on GPU. After training our model, we generated captions for some testing images in Flickr 8k Data and some pictures from the internet. Some of the captions were very good and, some were bad too. But in almost all images, our model was able to detect which objects are present in them. Some of the examples of generated captions are given below.

dog is running through the grass

man is climbing up rock face

References :

- Train your own image classifier with Inception in TensorFlow
- A picture is worth a thousand (coherent) words: building a natural description of images
- How to Develop a Deep Learning Photo Caption Generator from Scratch
- Deep Visual-Semantic Alignments for Generating Image Descriptions
- Flickr 8k Data