CAPTIONING OF IMAGES

5.1 BI DIRECTIONAL LSTMS

LSTM, or Long Short Term Memory, is a modified architecture aimed to remove the problem of Gradient Vanishing, or loss of information, in Recurrent Neural Networks (RNNs).

It includes copying the principal intermittent layer in the system so that there are presently two layers next to each other, at that point giving the input grouping to the main layer and giving a turned around duplicate of the input information to the second. [13]

The cell state in the architecture is somewhat similar to a transport line. It runs straight down the whole chain, with just some minor direct connections. The LSTM has the capacity to expel or add data to the cell state, deliberately managed by structures called Gates. Gates are an approach to alternatively let data through. They are made out of a sigmoid neural net layer and a pointwise augmentation task.

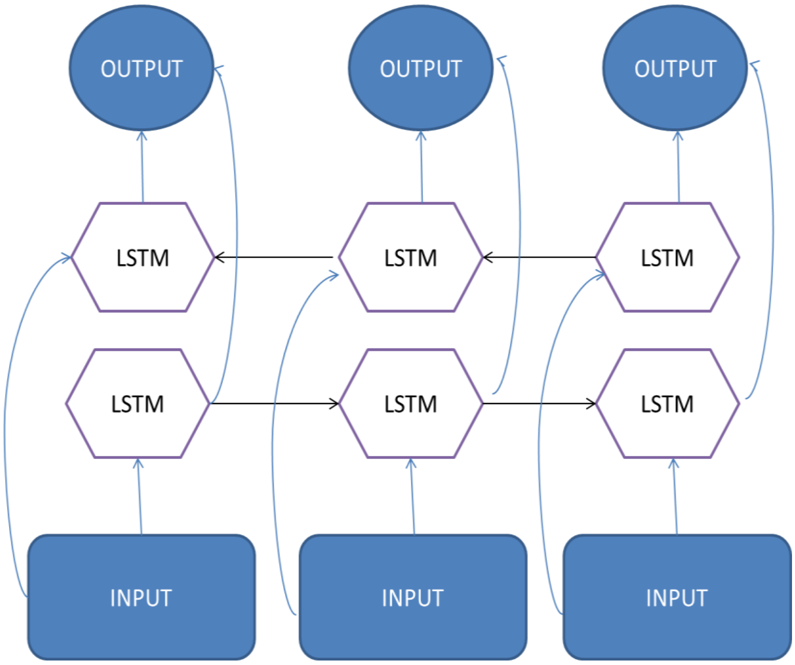


Fig 8: Representation of a Bi Directional LSTM

Mathematically, it can represent it as,

Input gate:

$X^i = \sum^n_{x=1} W*k_x + \sum^n_{x=1} W*l_x^{t-1} + \sum^n_{x=1} W*m_x^{t-1}$

$l_x = f(X^i)$

…(21)

Forget gates:

$Fr^i = \sum^n_{x=1} W*k_x + \sum^n_{x=1} W*l_x^{t-1} + \sum^n_{x=1} W*m_x^{t-1}$

…(22)

Cell Input:

$C = \sum^n_{x=1} W*k_x + \sum^n_{x=1} W*l_x$

$m_x = l^t*m^{t-1} + l^t*h(X^i)$

…(23)

Cell Output:

$O = \sum^n_{x=1} W^c*p_x^t + \sum^n_{x=1} W^c*p_x^{t+1}$

$l_t = l_t*h(m_t)$

…(24)

Output gate:

$p_x = f'(X^i) \sum^n_{x=1} h(m_t)O$

…(25)

This project uses Bi Directional LSTMs for building the model for the captioning of the image. Due to its bi direction nature, the input is fed in two directions, basically forward and backward, hence giving more accurate results.

5.2 INCEPTION V3

Inception v3 is a broadly utilized picture acknowledgment model that has been appeared to accomplish more noteworthy than 78.1% precision on the various modelling dataset. The model is the perfection of numerous thoughts created by different specialists throughout the years.

The model itself is comprised of symmetric and lopsided structures, including convolution layers, pooling, maximum pooling, fallouts, and completely associated layers.

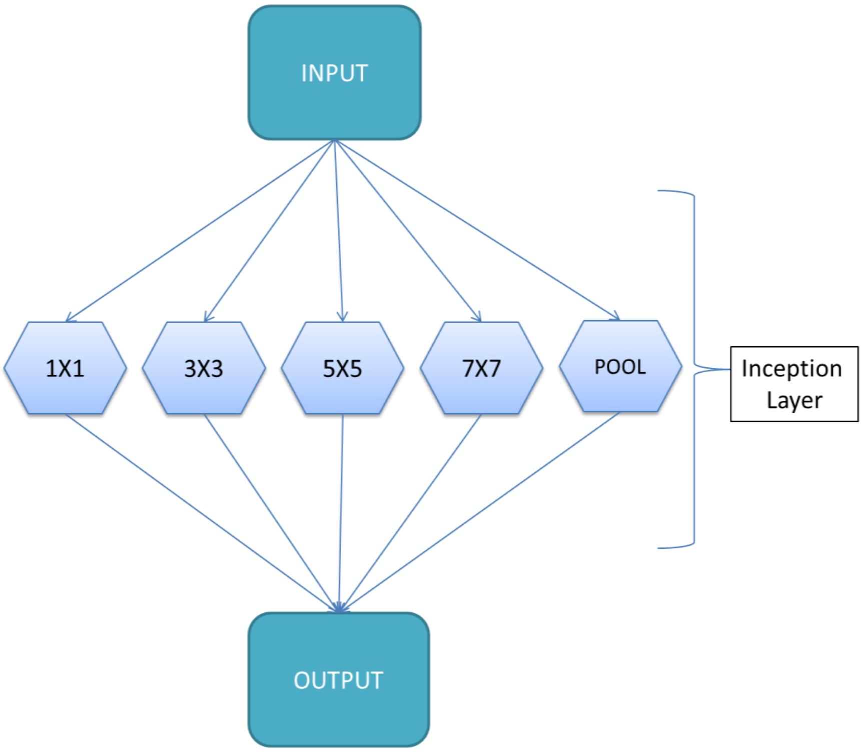


Fig 9: Inception Model Depiction

The basic idea is that at each layer, it can pick either having a pooling operation or a convolution of 1 by1 or 3 by 3 and so on. Inception Model combines all of this, i.e. instead of having a single layer; it has a pooling layer with a 1 by1 or 5 by 5 convolution layer and so on. The output from all these are simply concatenated, giving more accurate results. [15]

Let us say the input to the inception layer, made up of five sub layers, is I,

$$O_1 = F_1(I)$$
$$O_2 = F_2(I)$$
$$O_3 = F_3(I)$$
$$O_4 = F_4(I)$$
$$O_5 = F_5(I)$$

…(26)

Thus the final output of the inception layer will be

$$O = O_1 + O_2 + O_3 + O_4 + O_5 \quad \ldots(27)$$

The project uses the third version of the Inception series model, which is a bit more sophisticated and have a better training. Inception V3, due to its better classification capacity, gave quite efficient results in the captioning of the images.

5.3 CAPTIONING MODEL

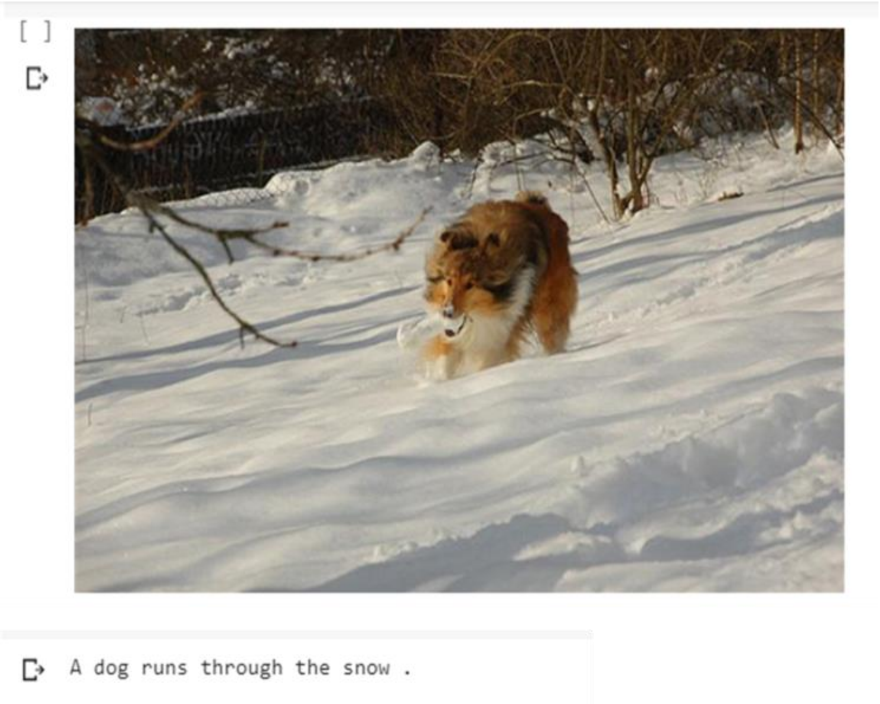
The project proposes an efficient captioning model for the image captioning. The model was made up using Bi Directional LSTMs and Inception V3 architecture, giving a firm base for the model to train and give good results. [14]

- Dataset: The model is trained using the Flicker8k dataset of images. With a total of 8000 images, it was standardly broken into 1000, 1000 and 6000 for testing, validation and training respectively.

- Architecture: The captioning model started with an Inception V3, combined with a dense layer, to encode and extract various features of the image. Rectified Linear Unit is used as an activation function in the proposed model.

  Bi directional LSTM is used for further decoding, and converting the vector image into a sequence sentence format. Dropouts are also introduced so to prevent the model from the problem of overfitting.

  A layer of neural network implementing Softmax is also introduced just before the output layer. Basically the Softmax layer finds out the various probabilities of different outcome solutions, the winner of which id fed into the output layer. All in all ten layers are used from the input to output layer, combining dropout layers, dense layers, and the lstm layer.



```
A dog runs through the snow .
```

Fig 10: (Upper) Input Image; (Lower) Caption predicted by the model

- Results: It was observed that with more epochs, or iterations, in the training of the model, the accuracy increased effectively, going quite close to fifty percent. Subsequently the loss also decreased considerably.

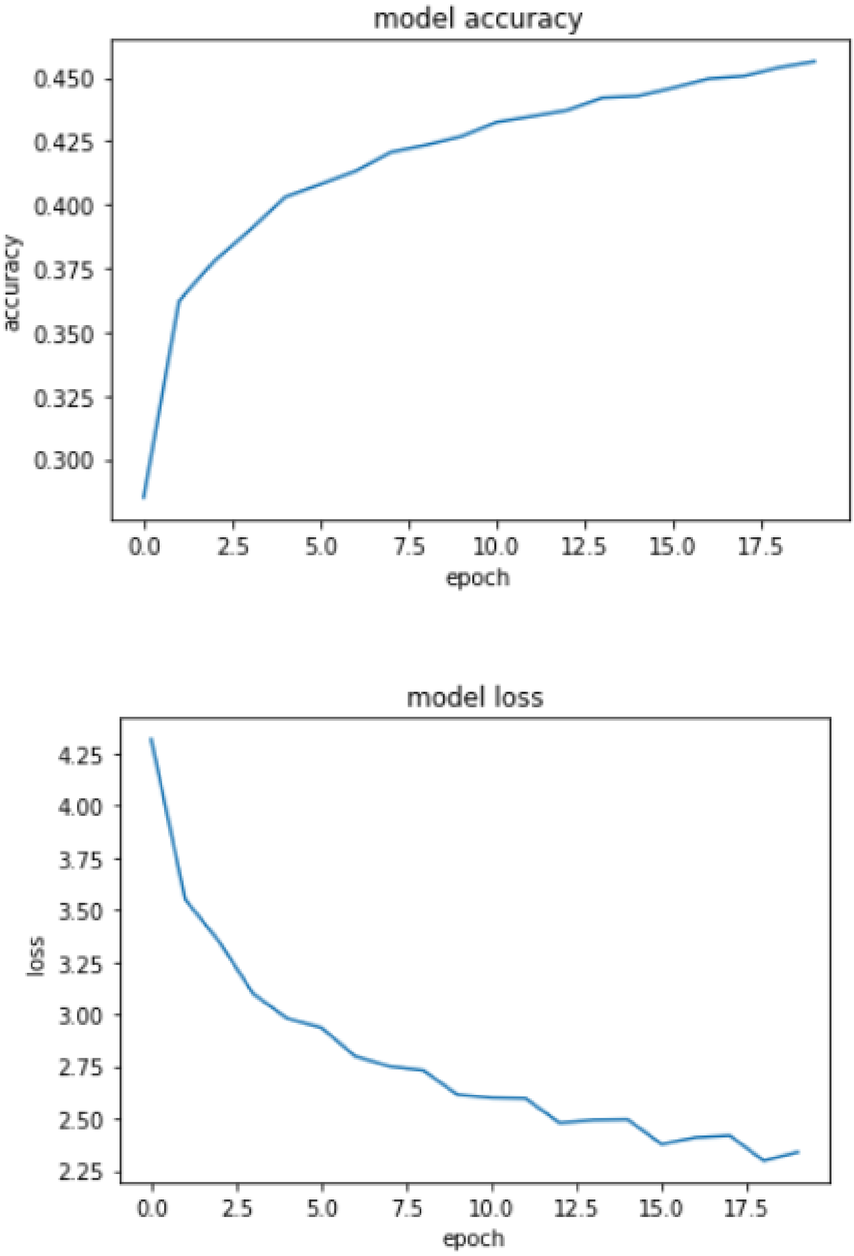


Fig 11: (Upper) Accuracy Graph against number of epochs; (Lower) Loss Graph against number of epochs

Thus the project proposes a model, using bi directional lstm and inception v3 neural networks for captioning of an image.