## Natural Language Processing Metrics - A Text Based Metrics For Image Captioning

Nitya Sree.P,

*Department of computer science and engineering,*

*Vellore Institute of Technology,*

Chennai, India

[nityasree.p2017@vitstudent.ac.in](mailto:nityasree.p2017@vitstudent.ac.in)

## *Abstract*: In the past few years, the problem of generating descriptive sentences automatically for images has garnered a rising interest in natural language processing and computer vision research. Image captioning is a fundamental task which requires semantic understanding of images and the ability of generating description sentences with proper and correct structure. In this study, the authors propose a hybrid system employing the use of VCG16 model to generate vocabulary describing the images and a Long Short Term Memory (LSTM) to accurately structure meaningful sentences using the generated keywords. We showcase the efficiency of our proposed model using the Flickr8K datasets and show that their model gives accurate results compared utilising the Bleu metric. The Bleu metric is an algorithm for evaluating the performance of a machine translation system by grading the quality of text translated from one natural language to another.

***Keywords:*** Deep Learning, Image Captioning, LongShort Term Memory (LSTM), Graphics Processing Unit(GPU), Tensor Processing Unit (TPU)

## ****INTRODUCTION :****

Caption generation is an interesting artificial intelligence problem where a descriptive sentence is generated for a given image. It involves the dual techniques from computer vision to understand the content of the image and a language model from the field of natural language processing to turn the understanding of the image into words in the right order. Image captioning has various applications such as recommendations in editing applications, usage in virtual assistants, for image indexing, for visually impaired persons, for social media, and several other natural language processing applications. Recently, deep learning methods have achieved state-ofthe-art results on examples of this problem. It has been demonstrated that deep learning models are able to achieve optimum results in the field of caption generation problems. Instead of requiring complex data preparation or a pipeline of specifically designed models, a single end-to-end model can be defined to predict a caption, given a photo. In order to evaluate our model, we measure its performance on the Flickr8K dataset using the BLEU standard metric. These results show that our proposed model performs better than standard models regarding image captioning in performance evaluation.

1. **RELATED WORK:**

The image captioning problem and its proposed solutions have existed since the advent of the Internet and its widespread adoption as a medium to share images. Numerous algorithms and techniques have been put forward by researchers from different perspectives. Krizhevsky et al. [1] implemented a neural network using non-saturating neurons and a very efficient a unique method GPU implementation of the convolution function. By employing a regularization method called dropout, they succeeded in reducing overfitting. Their neural network consisted of maxpoolinglayers and a final 1000-way softmax. Deng et al. [2] introduced a new database which they called ImageNet, an extensive collection of images built using the core of the WordNet structure. ImageNet organized the different classes of images in a densely populated semantic hierarchy. Karpathy and FeiFei [3] made use of datasets of images and their sentence descriptions to learn about the inner correspondences visual data and language. Their work described a Multimodal Recurrent Neural Network architecture that utilises the inferred co-linear arrangement of features in order to learn how to generate novel descriptions of images. Yang et al. [4] proposed a system for the automatic generation of a natural language description of an image, which will help immensely in furthering image understanding. The proposed multimodel neural network method, consisting of object detection and localization modules, is very similar to the human visual system which is able to learns how to describe the content of images automatically. In order to address the problem of LSTM units being complex and inherently sequential across time, Aneja et al. [5] proposed a convolutional network model for machine translation and conditional image generation. Pan et. al [6] experimented extensively with multiple network architectures on large datasets consisting of varying content styles, and proposed a unique model showing noteworthy improvement on captioning accuracy over the previously proposed model

1. **PROPOSED SYSTEM:**

The basic working of the project is that the features are extracted from the images using pre-trained VGG16 model and then fed to the LSTM model along with the captions to train. The trained model is then capable of generating captions for any images that are fed to it.The dataset used here is the [FLICKR 8K](https://forms.illinois.edu/sec/1713398) which consists of around 8091 images along with 5 captions for each images.

The following are the Dependencies needed for the project:Keras, Tensorflow GPU, Pre-trained VGG-16 weights, NLTK, Matplotlib

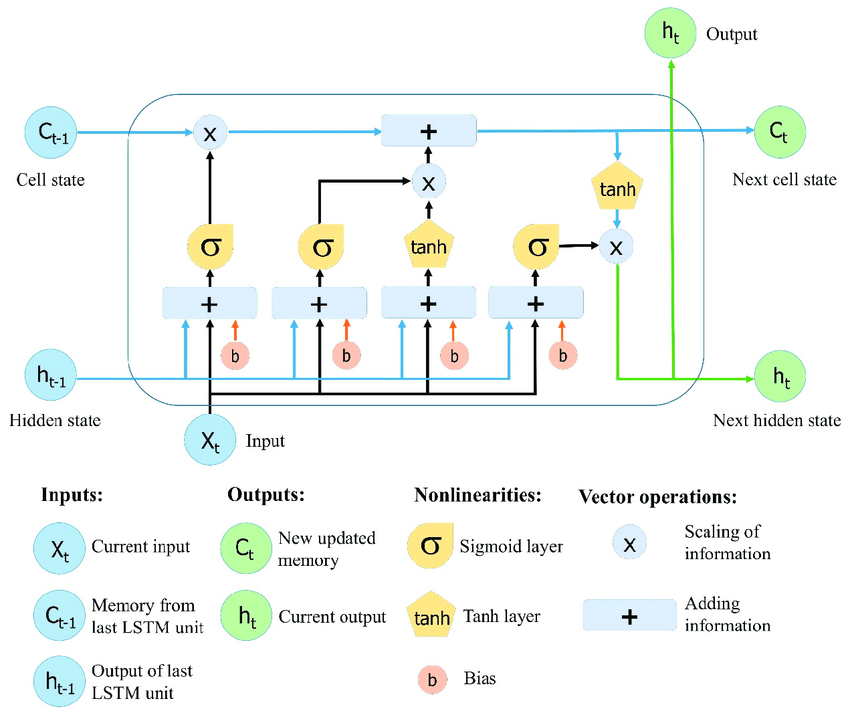


Fig.1. LSTM

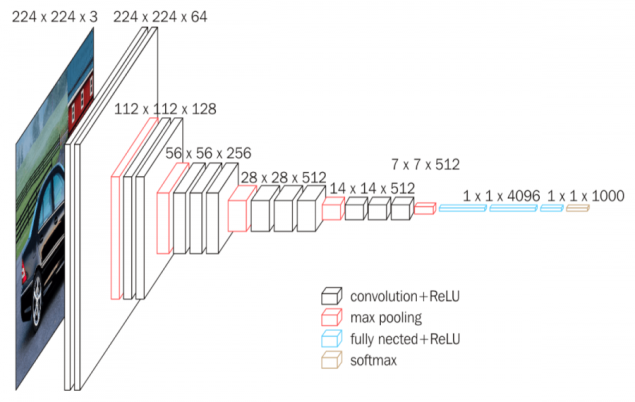


Fig.2. VCG 16

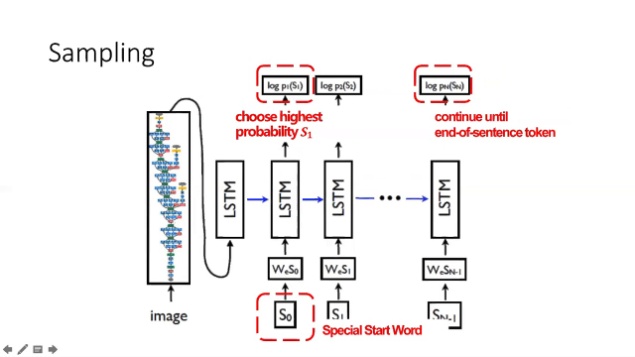


Fig.3. system Design

### IMPLEMENTATION:

### a)Data preprocessing:

### Importing the required libraries: all the required libraries like keras, pandas, NLTK, matplotlib, tensorflow are imported

**Configuring:** now we are configuring the GPU memory to be used for training purposes. We set it such that 95% of the available memory of the GPU are used

## Importing dataset : we now import the flicker image dataset and its respective captions

## understanding the dataset: Plotting few images and their captions from the dataset to understand more about the dataset

## Cleaning the captions for further processing:The caption dataset contains punctuations, singular words and numerical values that need to be cleaned before it is fed to the model because uncleaned dataset will not create good captionsfor the images

## Analysing caption in the dataset: Plotting the top 50 words that appear in the cleaned dataset

## Adding start and end sequence tokens for each captions:

Start and end sequence has to be added to the tokens so that it is easier to identify the captions for the images as each of them are of different length

## b)Extract features:

## We Load VGG16 model and weights to extract features from the images

The last layer of the VGG-16 is excluded here because we are are just using it for extracting the features rather than using for object classification

Here the features are extracted from all the images in the dataset. VGG-16 model gives out 4096 features from the input image of size 224 \* 224

## Plotting similar images from the dataset:

For this we have to first create a cluster and find which images belong together. Hence PCA is used to reduce the dimensions of the features which we got from VGG-16 festure extraction from **4096**to**2**

First the clusters are plotted and few examples are taken from the bunch for displaying

## c) merging dataset

## Merging the data: merging the images and the captions for training

## Tokenizing:tokenizing the captions for further processing

As the model can't take texts as an input, they need to converted into vectors.

## Splitting the data: The data is split as training and test data

## Padding: Finding the max length of the caption for padding

## Processing: Processing the captions and images as per the required shape by the model

**d) Building model:**

**Building the LSTM model:**LSTM stands for Long Short Term Memory is a kind of recurrent neural network. Long Short Term Memory can by default retain the information for long period of time in memory. Long Short Term Memory is used for processing, predicting and classifying based on time series data. A recurrent neural network is ideal for sequences, lists, and other language processing problems it allows information to be passed from one step of a network to another. A LSTM is capable of learning long-term dependencies and work incredibly well on a large variety of problems. While RNNs have difficulty with long-term dependencies, LSTMs are explicitly designed to avoid the long-term dependency problem.

## Training the LSTM model: train the lstm model for epoch times

## Generating captions on a small set of images:

After the model finishes training we can test out its performance on the some of the test images to figure out if the generated captions are good enough. If the generated captions are good enough we can generate the captions for the whole test dataset.

## e) Evaluating the model performance:

After the model is trained we have to test the models prediction capabilities on test dataset. Traditional accuracy metric can't be used on predictions. For text evaluations we have a metric called as [BLEU Score](https://machinelearningmastery.com/calculate-bleu-score-for-text-python/). BLEU stands for Bilingual Evaluation Understudy, it is a score for comparing a candidate text to one or more reference text.

## Generating captions for the whole test data and finding BLEU score

We can check out some of the generated caption's quality as good and bad caption based on BLEU score. Some times due to the complex nature of the images the generated captions are not acceptable.

**V. RESULTS AND DISCUSSION:**

The image captioning model was implemented and we were able to generate moderately comparable captions with compared to human generated captions. Using PCA we plot similar images from the dataset

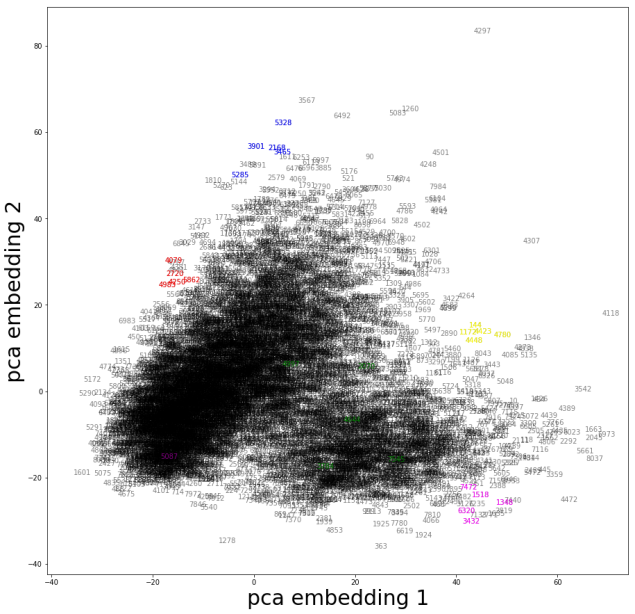


Fig.3.

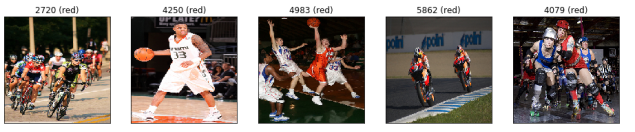
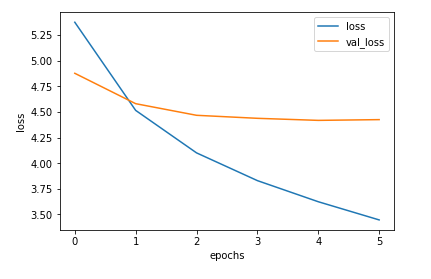




Fig.4.pca embedding



Fig,5.

Fig: loss vs epoch and val\_lossvs epoch for lstm model



Fig.6. bad caption

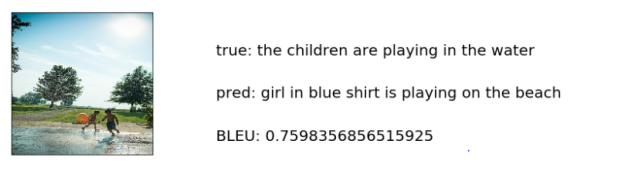


Fig.7..good caption

## CONCLUSION:

The model has been successfully trained to generate the captions as expected for the images. The caption generation has constantly been improved by fine tuning the model with different hyper parameter. Higher BLEU score indicates that the generated captions are very similar to those of the actual caption present on the images.

**VII. REFERENCES:**

[1]Alex Krizhevsky, IlyaSutskever, and Geoffrey E. Hinton, ImageNet Classification with Deep Convolutional Neural Networks, [Online] Available: https://papers.nips.cc/paper/4824-imagenetclassificationwith-deep-convolutional-neural-networks.pdf [2]Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li and Li Fei-Fei, ImageNet: A Large-Scale Hierarchical Image Database [3]Andrej Karpathy, Li Fei-Fei, Deep VisualSemantic Alignments for Generating Image Descriptions, [Online] Available: https://cs.stanford.edu/people/karpathy/cvpr2015.pdf [4]Zhongliang Yang, Yu-Jin Zhang, SadaqaturRehman, Yongfeng Huang, Image Captioning with Object Detection and Localization, [Online] Available: https://arxiv.org/ftp/arxiv/papers/1706/1706.02430.pdf [5]JyotiAneja, AdityaDeshpande, Alexander Schwing, Convolutional Image Captioning, [Online] Available: https://arxiv.org/pdf/1711.09151.pdf [6]Jia-Yu Pan, Hyung-Jeong Yang, Pinar Duygulu, Automatic Image Captioning, Conference: Conference: Multimedia and Expo, 2004. ICME ’04. 2004 IEEE International Conference on, Volume: 3 [7]OriolVinyals, Alexander Toshev, SamyBengio, DumitruErhan, Show and Tell: A Neural Image Caption Generator, [Online] Available: https://arxiv.org/pdf/1411.4555.pdf [8]Kelvin Xu, Jimmy Lei Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, RuslanSalakhutdinov, Richard S. Zemel, YoshuaBengio, Show, Attend and Tell: Neural Image Caption Generation with Visual Attention, [ Online ] Available: https://arxiv.org/pdf/1502.03044.pdf [9] M. Hodosh, P. Young and J. Hockenmaier (2013) ”Framing Image Description as a Ranking Task: Data, Models and Evaluation Metrics”, Journal of Artificial Intelligence Research, Volume 47, pages 853-899