IMAGE CAPTIONING USING CNN AND LSTM

G Sai Nagarjun Student, Department of Computer Science, Narayana Engineering College Gudur
K Harindranath Reddy Student, Department of Computer Science, Narayana Engineering College Gudur
Dr.P.Venkateswara Rao Professor, Department of Computer Science, Narayana Engineering College Gudur
K Rohith Karthikeya Student, Department of Computer Science, Narayana Engineering College Gudur

Abstract
The problem of automatically generating descriptive sentences for images has sparked renewed interest in natural language processing and computer vision research in recent years. Image captioning is a key task that necessitates a semantic comprehension of images as well as the capacity to generate accurate and precise description phrases. In this research, the authors propose a hybrid system that employs a multilayer Convolutional Neural Network (CNN) to get vocabulary for describing images and an extended Short Term Memory (LSTM) to accurately structure meaningful sentences using the generated keywords. The target image is compared to a large dataset of images by the convolutional neural network. The trained captions are then used to generate an accurate description of the training images. We demonstrate the efficiency of our proposed model using the Flickr8K and Flickr30K datasets, as well as generated captions for sample images and compared the different feature extraction and encoder models to determine which model provides the best accuracy and produces the desired results. The effectiveness of the proposed model outperforms previous benchmark models when evaluated using standard evaluation metrics.
Keywords: Convolutional Neural Network, Long-Short Term Memory, Image Captioning.

Introduction
Caption generation may be a stimulating AI problemwhere 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 world of tongue processing to point out the understanding of the image into words within the proper order. Image captioning has various applications like recommendations in editing applications, usage in virtual assistants, for image indexing, for visually impaired persons using text-to-speech through real time feedback about encompassing the situation over a camera feed, improving social medical leisure with the aid of reorganizing the captions for photographs in social feed along with messages tospeech. Recently, deep learning methods have achieved state-of the-art results on samples of this problem.[1][2] It's been demonstrated that deep learning models are able to achieve optimum results in the world of captiongeneration problems rather than requiring complex data preparation or a pipeline of specifically designed models, one end-to-end model are often defined to predict a caption. Every day we see a lot of photographs in the surroundings, on social media and in the newspapers. Humans are able to recognize photographs themselves only[3]. We humans can pick out the photographs without their designated captions but on the other hand machines need images to get trained first then it'd generate the photograph caption automatically. Social media like Instagram, Facebook etc can generate captions routinely from images.

The principal goal of this research paper is to get a little bit of expertise in deep learning strategies. We use two strategies specially CNN and LSTM for image classification. These results show that our proposed model performs better than standard models regarding image captioning in performance evaluation[4].

Related work
The image captioning problem and its proposed solutions have existed since the arrival of the web

and its widespread adoption as a medium to share images[5][6]. Numerous algorithms and techniques are suggests by researchers from different perspectives. Krizhevsky et al. Implemented a neural network using non-saturating neurons and a really efficient a singular method GPU implementation of the convolution function. By employing a regularization method called dropout, they succeeded in reducing overfitting[7]. Their neural network consisted of maxpooling layers and a final 1000-way softmax. Deng et al. Introduced a replacement database which they called ImageNet, an in depth collection of images built using the core of the WordNet structure. ImageNet organized the various classes of images during a densely populated semantic hierarchy. Karpathy and FeiFei made use of datasets of images and their sentence descriptions to find out about the inner correspondences visual data and language. Their work described a Multimodal Recurrent Neural specification that utilizes the inferred co-linear arrangement of features so as to find out the way to generate novel descriptions of images. Yang et al. proposed a system for the automated generation of a tongue description of a picture, which can help immensely in furthering image understanding. The proposed multimodel neural network method[8][9], consisting of object detection and localization modules, is extremely almost like the human sensory system which is in a position to learns the way to describe the content of images automatically. so as to deal with the matter of LSTM units being complex and inherently sequential across time, Aneja et al. proposed a convolutional network model for MT and conditional image generation. Pan et. al experimented extensively with multiple network architectures on large datasets consisting of varying content styles, and proposed a singular model showing noteworthy improvement on captioning accuracy over the previously proposed models. Vinyals et al. presented a generative model consisting of a deep recurrent architecture that leverages MT and computer vision, wont to generate natural descriptions of a picture by ensuring highest probability of the generated sentence to accurately describe the target image. Xuetal. Introduced an attention based model that learned to describe the image regions automatically. The model was trained using standard back propagation techniques by maximizing a variable lower bound. The model was able to automatically learn identify object boundaries while at the same time generate an accurate descriptivesentence.

Methodology

Let us discuss the various techniques and algorithms for image captioning.Figure 2 shows the Architecture of CNN-LSTM-based image captioning system. The architecture involves two main modules[10][11]. The first one is image understanding module using CNN and the second one is text understanding module using LSTM. Each module is described in details in the following subsections.

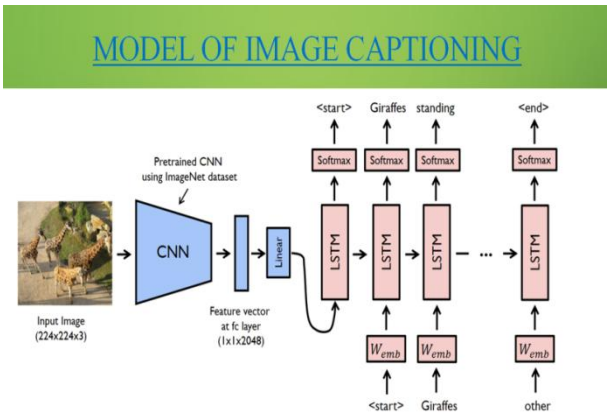

Fig 1: Image Captioning Model

Convolution Neural Network (CNN)

For image caption generation task, CNN is widely used because it has solved successfully for image annotation problems with high accuracy. Convolutional Neural systems are specific important neural

systems that can produce information that has an information shape,for example, a 2D lattice and CNN is valuable for working with pictures. It examines pictures from left corner to the right corner and through to extricate significant highlights from the picture, and consolidates the element to characterize pictures. It can deal with interpreted, pivoted, scaled, and modified pictures. The Convolutional neural system is a profound learning calculation that takes in the info picture, allocates significance to various components/protests in the picture, and recognizes it from each other.

Architecture of CNN

A pure rustic neural network, in whatever location all neurons in a single layer merge with all of the neurons in the subsequent layer is inefficient in regards to analyzing large pictures and video.

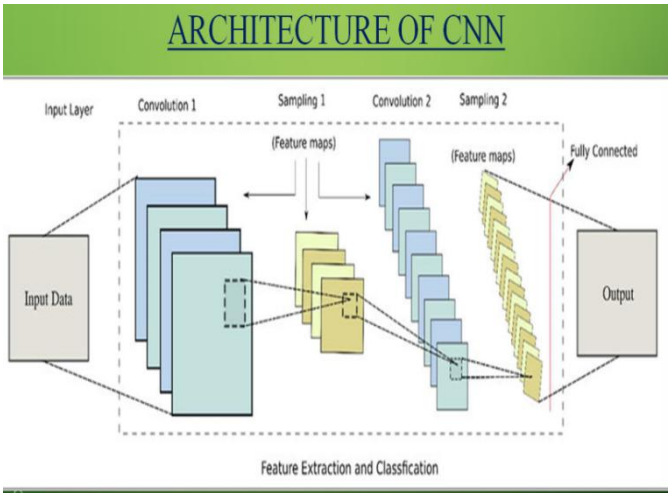

Fig 2: Architecture of CNN

How does CNN work?

As we have discussed previously, a fully connected neural network where the input in the preceding layers is connected to every input in the following layers is convenient for the task at hand, along such lines, according to CNN, the neurons in a cell may be connected with a specific cell area before it, rather than all the neurons in a totally similar way.

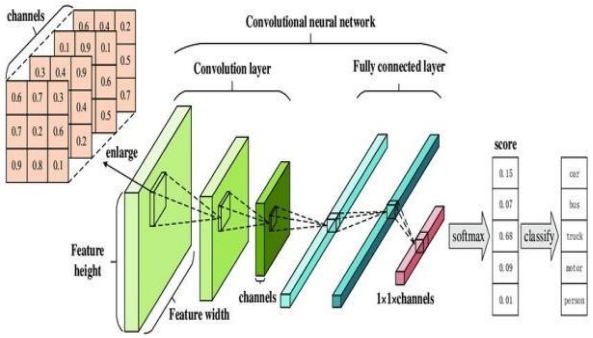

Fig 3: Working of CNN

This helps in reducing the complexity of the neural network and acquiring less computing power. When we generally compare two images we check the pixel values of each pixel. This technique only helps us to compare two identical images only but when we keep different images to compare the comparison fails. In CNN image comparison takes place piece by piece.

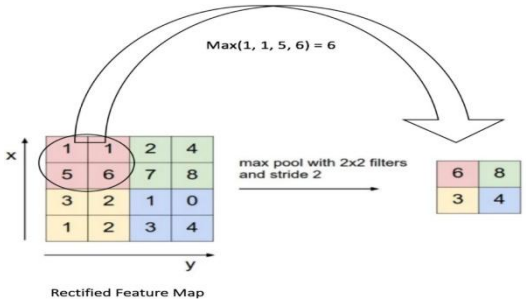

Fig 4: Feature map of CNN picture

The main reason behind using CNN algorithm is that, this is the only algorithm which takes pictures as an input and on the basis of input pictures drawing the feature map, ie.classifying each pixel on the basis on similarity and differences. The CNN classifies the pixels and a matrix is created, which is known as feature map. Feature map is a collection of similar pixels placed in a separate category. These matrices play an important role in finding the essence of the thing in the input picture.
There are total 3 types of layers in CNN model-

1. Convolutional
2. Pooling
3. Fullyconnected

In the first layer, the input image is read through the CNN, and on that foundation a feature map is made. From that feature map, it serves as an input to the following layers, i.e for the Pooling layer. In the pooling layer, the feature map is broken down into extra simpler parts to carefully examine the context of the picture. This layer makes the feature map more dense so as to discover the most critical information about the picture.
The 1st and 2nd layers i.e Convolutional and Pooling they're practised so many times, depending on the picture as to get the densed information about the picture. The extra dense feature map is created because of these two layers. And thisdensed feature map is utilized by the last layer i.e FullyConnected.
This layer performs classification. It sorts the pixels with respect to similarity and differences. Classification is done upto exceptional limit so as to get the essence of the picture, help in identifying the objects, persons, things,etc.

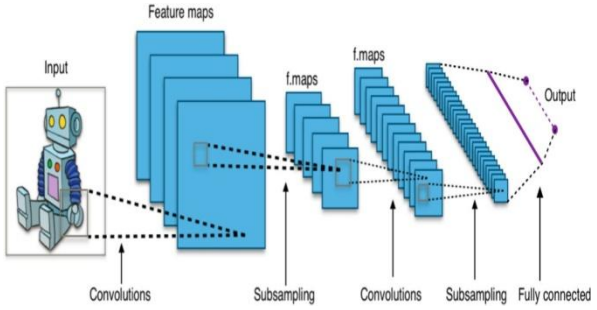

Fig 5: Layers of scanned picture

Vgg16
It is a pre-trained model on ImageNet dataset based on Visual Geometry Group (VGG) OxfordNet 16-layer CNN (Rahul and Aayush, 2018; Lakshminarasimhan et al., 2018). The VGG16 neural network is used for image classification. Output of VGG16 is probability of individual classes that the classification system has to classify. We remove the last layer of the VGG16 and use the output from second last layer as feature parameters for each image. We extract 4096 parameters for each

image, which are further processed by a Dense layer to produce a 256 element representation of an image (Micah et al., 2013).

Recurrent Neural Networks (RNN)

RNNs are a part of a deep learning set of rules which are performed to deal with a number of complicated or complex computer tasks like item classification & speech recognition. RNNs are performed to address an array of activities that arise in series, with the information of every situation based completely on statistics from preceding situations.This RNN can be used to carry out plenty of real life problems like inventory forecasting & reinforce speech recognition. Yet, RNNs are not used to solve real life problems & that is because of the Vanishing Gradient problem.

This vanishing gradient problem is the main cause which makes the working of RNNs challenging. In general, the engineering of RNNs is made such that it stores the data for some short period of time and stores some array of data. It's not possible for RNNs to remember all the data values and a long period of time. RNNs can only store some of the data for a small period of time.To solve this problem, we will be using Long short-term memory (LSTM), which is a subset of RNNs. LSTM are basically constructed toovercome the problem of Vanishing Gradients.

Long Short Term Memory (LSTM)

LSTM can maintain information in memory for long periods of time and retrieve sequential information through time (Yang et al., 2018). The text understanding part produces words or phrases based on the word embedding vector of previous part. The language generation model is trained to predict each word in the caption after it has seen both image and all previous words. For any given sentence in Myanmar corpus we add two extra symbols for start word and stop word which designates the start and end of the sentence. Whenever stop word is found it halts generating caption and it denotes end of the sentence. The understanding of LSTM gates to hold the data for a period of time offers benefit to the LSTM over the RNNs.

A**rchitecture of LSTM**

The architecture of LSTM is very simple, it consists of 3 major gates, which store the data for a longer period of time and help in solving the difficulties which RNNs couldn't solve.

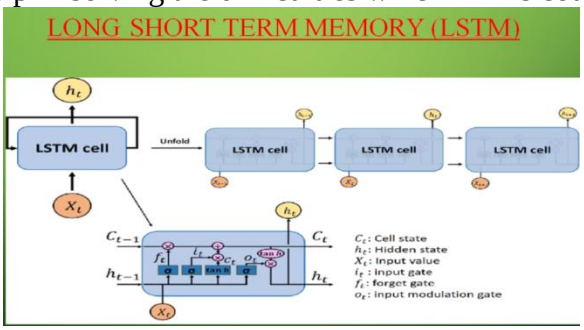

Fig 6: LSTM Architecture

The 3 major gates of the LSTM covers are:

- Forget gate — the main work of the forget gate is to filter the data, i.e. to delete all that data whichis not needed in the future to solve a particular task. This gate is responsible for the overall performance of the LSTM, it optimizes the data.
- Input gate — the starting of LSTM starts from this gate, i.e. input gate. This gate takes input from the user and supplies the input data to othergates.
- Output gate — This gate is responsible for showcasing the desired result in a propermanner.

Uses of LSTM

LSTMs are profoundly and mostly used for variety deep learning duties that largely encompass forecasting of the data depending upon the preceding data. The 2 remarkable illustrations cover text prediction and stock marketprediction.

Modules

For our research purpose, we have downloaded the data set which consists of following files:

- Flickr8k_Datasets**:** This file contains all the pictures for which we have to first train our model. It contains 8091images.
- Flickr8k_texts: This folder contains text files & pre-formed captions for thepictures.
- Description.txt: This is the file which will contain the picture names & their related captions laterpreprocessing.
- Feature.p**:** This file binds the picture and their related captions that are extracted from the Xception, which is a pre-trained CNNmodel.
- Tokenizers.p**:** This file contains an expression which we call tokens, and these tokens are generalizedwith the indexvalue.
- Models.png**:** Diagrammatic representation of extension of the CNN-LSTMmodel.
- Testing_captions_generator.py**:** This is the Python file which is used in generating the captions of thepictures.
- Training_captions_generator.ipynb**:** This is basically a Jupyter notebook, which is in short a web based application. We use this to train our model & on that basis achieving captions to our inputpictures.

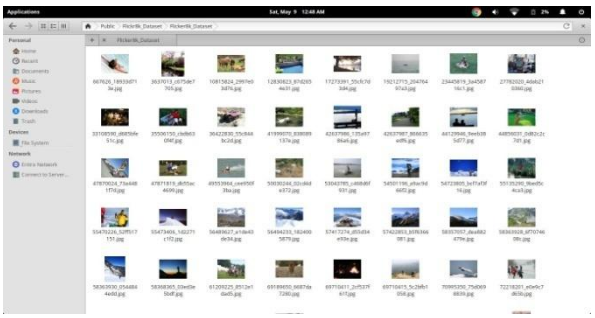

Fig 7: Flickr_Dataset

Results

The caption generated for the input imagesare shown below.

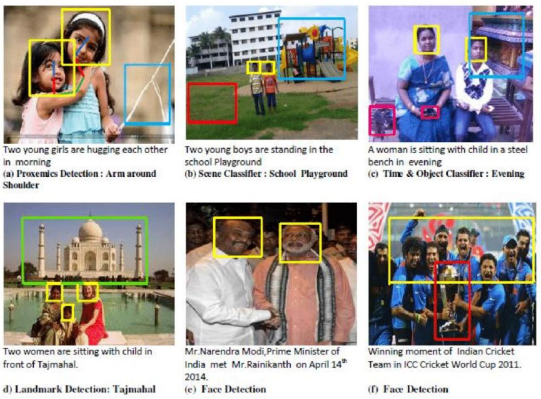

Fig 8: Output

In figure 9(a), the model accurately generated the major features in the images such as “Two young girls are hugging each other in morning.” And the relationship between these features of images also describes accurately. Figure 9(b) also generates the caption as “Two young boys are standing in the school playground.” which is grammatically correct. Similarly, all the images are captioned automatically without any human interference.

Conclusion

The CNN-LSTM model was built on the idea of generating the captions for the input pictures. This model can be used for a variety of applications. In this, we studied about the CNN model, RNN models, LSTM models, and in the end we validated that the model is generating captions for the input pictures.

References

1.Alex Krizhevsky, IlyaSutskever, and Geoffrey E. Hinton, Image Net Classification withDeep 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. ZhongliangYang,Yu-JinZhang,SadaqaturRehman, Yongfeng Huang, Image Captioning with Object Detection and Localization, [Online] Available: https://arxiv.org/ftp/arxiv/papers/1706/1706.02430.p df
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.Sucharita, V., Venkateswara Rao, P., Bhattacharyya, D., Kim, T.-H. Classification of penaeid prawn species using radial basis probabilistic neural networks and support vector machines International Journal of Bio-Science and Bio-Technology, 2016, 8(1), pp. 255–262
8.V. Sucharita, S. Jyothi ,D.M. MamathaA Comparative Study on Various Edge Detection Techniques used for the Identification of Penaeid Prawn Species ,International Journal of Computer Applications (0975 – 8887) Volume 78 – No.6, September 2013
9. OriolVinyals, Alexander Toshev, SamyBengio, DumitruErhan, Show and Tell: A Neural Image Caption Generator, [Online] Available: https://arxiv.org/pdf/1411.4555.pdf
10. Kelvin Xu, Jimmy Lei Ba, Ryan Kiros, Kyunghyun cho,AaronCourville,RuslanSalakhutdinov,RichardS. Zemel, YoshuaBengio, Show, Attend and Tell: Neural Image Caption Generation with Visual Attention, [Online] Available: https://arxiv.org/pdf/1502.03044.pdf
11. S. Jyothi, V. Sucharita, D.M. Mamatha " Survey on Computer Vision and Image Analysis based Techniques in Aquaculture" CIIT International Journal of Digital Image Processing, 2013