**Automatic Image Captioning Using Convolutional Neural Network And Deep Learning**

**Akshay RajendraGhatole, Mihir Singh, Mohan Dongara and Himakar Sandiri**

**Group No - 17**

**ABSTRACT**

Image Captioning is the process of generating a textual description for given images. It has been a very important and fundamental task in the Deep Learning domain. The big challenge lies in obtaining semantic information from images and translating that into the human language using machines. The interaction of computer vision and natural language processing further increases the complexity of image captioning. Automatically describing the content of an image is a fundamental problem in artificial intelligence that connects computer vision and natural language processing.When a recurrent neural network language model is used for caption generation, the image information can be fed to the neural network either by directly incorporating it in the RNN. In this paper,we present the multimodal neural network that uses feature vectors obtained using both RNN and CNN, so consequently, we will have two inputs. One is the image we need to describe, a feed to the CNN, and the second is the words in the text sequence produced till now as a sequence as the input to the RNN.The model is trained to maximize the likelihood of the target description sentence given the training image. The dataset used here is Flickr8k\_Dataset for classifying image with their captions.Experiments on several datasets show the accuracy of the model and the fluency of the language it learns solely from image descriptions. Forgetting accuracy,like summarization, language translations, or captioning we use a Metrics called the BLEU score.In addition, the proposed method is competitive with many state-of-the-art methods in terms of standard evaluation metrics.

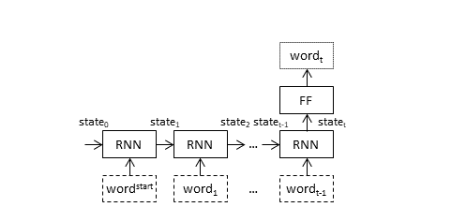
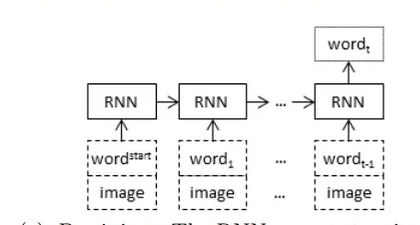
**INTRODUCTION**

In Today’s World , being able to automatically describe the content of an image using a common language is a challenging and tedious task, but has a great impact on the internet or web to help people understand the context of the image.A large amount of information is stored in an image. Everyday huge image data is generated on social media and observatories.Image caption generation is the task of generating a natural language description of the content of an image, also known as a caption.This task is significantly harder, for example, than the well-studied image classification or object recognition tasks, which have been a main focus in the computer vision community [2]. A description must capture not only the objects contained in an image, but it also must express how these objects relate to each other. Moreover, the above semantic knowledge has to be expressed in a natural language like English, which means that a language model is needed in addition to visual understanding.

Most previous attempts have proposed to stitch together existing solutions of the above sub-problems, in order to go from an image to its description.Recent work on image captioning has answered this question in different ways, suggesting different views of the relationship between image and text in the caption generation task[1]. To our knowledge, however, these different models and architectures have not been systematically compared. Yet, the question of where image information should feature in captioning is at the heart of a broader set of questions concerning how language can be grounded in perceptual information, questions which have been addressed by cognitive scientists[1].

The two most common approaches to image captioning are top-down and bottom-up. Unlike the top-down approach, which begins with an image and then converts it to words, the bottom-up approach begins with words that describe the various aspects that are combined[3].Auto Image Captioning plays out its capacity in a succession of undertakings. The initial move towards understanding a picture starts with the extraction of the picture with its relative encompassing the following stage, the connection between the distinguished articles has been recognized for additional assessment for example for objects book and table, the connection between two to be characterized as the book is on the table[2].

For the main undertaking for example for separating the highlights out of the picture we have utilized Convolutional Neural Network (CNN) in this task. It is essential to take note of that extracting feature alludes to eliminating the last softmax layer in most cases. For the subsequent part, which is to produce a printed depiction we will utilize Long Short Term Memory (LSTM)[3]. LSTMs are a special type of RNN which are used to avoid the long term dependency problems which often occurs in case of RNNs.Here, we propose to follow this elegant recipe,by using the encoder RNN with a deep convolution neural network (CNN). Over the last few years it has been convincingly shown that CNNs can produce a rich representation of the input image by embedding it to a fixed-length vector, such that this representation can be used for a variety of vision tasks [3]. Hence, it is natural to use a CNN as an image encoder, by first pre-training it for an image classification task and using the last hidden layer as an input to the RNN decoder that generates sentences shown in Figure 1 [4]. We call this model the Neural Image Caption.

**Figure 1 Neural Image Caption(Tanti,2018,p.2)**

Figure 1: How RNN-based neural language models work. Legend: RNN - recurrent neural network; FF - feed forward layer; wordi - the ith generated word in the text; wordstart - the START token which is an artificial word placed at the beginning of every sentence in order to still have a prefix when predicting the first word (likewise there is an END token to predict the end of a sentence). Note that state1 represents the prefix ‘wordstart’, state2 represents the prefix ‘wordstart word1’, etc. After processing a prefix, the RNN passes its final state to a feedforward layer which then predicts how likely each known word is to be the next word in the prefix[4].

**LITERATURE REVIEW**

In Paper [1] it was obviously indicated that the Deep learning has begun getting a great deal of consideration in most recent couple of years and a ton of headways have additionally been made in this field. This is very obvious too when we take a gander at the details. In 2015 just four effective articles were distributed however after that the ubiquity of the field developed dramatically and it tends to be seen from the way that 57 articles were distributed in 2017–2018.The problem of generating natural language descriptions from visual data has long been studied in computer vision, but mainly for video. This has led to complex systems composed of visual primitive recognizers combined with a structured formal language, e.g. And-Or Graphs or logic systems, which are further converted to natural language via rule-based systems. Such systems are heavily hand-designed, relatively brittle and have been demonstrated only on limited domains, e.g. traffic scenes or sports.

In Paper [2] done by Di Lu and Spencer Whitehead proposed that another assignment can be made for which picture depictions will be given as contribution to the framework. The paper likewise referenced that the Image Captioning which has been being used presently does not have the particular inspiration of elements that frames the fundamental structure of picture. In this paper, they additionally proposed the answer for this issue. The paper recommended that CNN-LSTM model should be prepared so it will be fit for producing inscription dependent on pictures spoke to it.The problem of still image description with natural text has gained interest more recently. Leveraging recent advances in recognition of objects, their attributes and locations, allows us to drive natural language generation systems, though these are limited in their expressivity. Farhadiet al. [2] use detections to infer a triplet of scene elements which is converted to text using templates. Similarly, Li et al[2]. start off with detections and piece together a final description using phrases containing detected objects and relationships. A more complex graph of detections beyond triplets is used by Kulkarni et al[2]., but with template-based text generation. More powerful language models based on language parsing have been used as well. The above approaches have been able to describe images in the wild, but they are heavily hand- designed and rigid when it comes to text generation.

In Paper [3] Julakanti also proposed a solution based on CNN-LSTM based architecture only. The model uses the CNN to extract the features of a given image, which later is fed into the RNN or LSTM model. Later the RNN or LSTM model describes the image in grammatically correct form that can describe what is going in the image.The paper also discussed the advantage of Image captioning model to visually impaired person. To help visually impaired people in society, image captioning can come out to be a helpful device if developed accurately.An overall methodology of this class is to examine the visual substance of the picture first and afterward make subtitles from the visual vectorized content utilizing a language model. These strategies can make new inscriptions for each picture that is semantically more exact than past approaches. Most epic subtitle age models utilize profound AI-based approaches. Subsequently, profound learning-based novel picture subtitle creating techniques are our essential premium in this writing Novel inscription age-based picture inscription models for the most part utilize visual space and profound learning-based models. Inscriptions can likewise be created from multimodal space. Profound learning-based picture inscribing strategies can likewise be arranged on learning procedures: Supervised learning, Reinforcement Learning, and Unsupervised Learning.

In Paper [4,5],a large body of work has addressed the problem of rank- ing descriptions for a given image Such approaches are based on the idea of co-embedding of images and text in the same vector space. For an image query, descriptions are retrieved which lie close to the image in the embedding space. Most closely, neural networks are used to co-embed images and sentences together [5] or even image crops and subsentences [4] but do not attempt to generate novel descriptions. In general, the above approaches cannot describe previously unseen compositions of objects, even though the individual objects might have been observed in the training data. Moreover, they avoid addressing the problem of evaluating how good a generated description is.

This paper considers all the previous research that has been done in this field as of now and is likewise influenced from those research. The majority of the works that we have contemplated utilizes CNN and RNN based design. An intriguing finding that we have from the past exploration done on this point is that "adding more layers to the model doesn't really implies that we will get more precision".In this work we combine deep convolutional nets for im- age classification with recurrent networks for sequence modeling, to create a single network that generates descriptions of images. The RNN is trained in the context of this single end-to-end network. The model is inspired by recent successes of sequence generation in machine translation [3,2], with the difference that instead of starting with a sentence, we provide an image processed by a convolutional net. The closest works are by Kiros et al. [5] who use a neural net, but a feedforward one, to predict the next word given the image and previous words. A recent work by Mao et al. [2] uses a recurrent NN for the same prediction task. This is very similar to the present proposal but there are a number of important differences: we use a more powerful RNN model, and provide the visual input to the RNN model directly, which makes it possible for the RNN to keep track of the objects that have been explained by the text. As a result of these seemingly insignificant differences, our system achieves substantially better results on the established benchmarks.

**PROPOSED APPROACH**

As of now examined in the theoretical, the fundamental point of this paper is to give subtitles to the picture in the real time. The dataset that is utilized for building this undertaking is Flickr8k dataset. In Flickr8k dataset each picture has 5 subtitles comparing to it. The images have been chosen from six different Flickr groups and does not contain well known locations or group.The following approaches will be used during the phases of the project:

**Data Cleaning:**

Getting the Image id from the Dataset and making a dictionary that will map pictures with the captions. The token.txt document has picture id and captions as sections and from this token.txt record just we will plan each picture with their individual inscriptions.The data will consists of Flickr8k\_Dataset and Flickr8k\_Text.Dataset consists of 8000 images of which 6000 will be used for training,1000 for test and 1000 for development.Text describes the train\_set and test\_set and contains token.txt with 5 captions for each image i.e total 40000 captions.For cleaning data removing of noise is necessary so that the text can be represented correctly.So Data cleaning will be used for clearing the noise by cleaning special characters such as hashtag,punctuations and numbers.

**PreProcessing**

Presently the images are used in such a way which need to be convert into machine understable. For this, We have utilized the transfer learning for example we utilize a pre-trained rendition that has been now taught on huge datasets and concentrate the capacities from these examples and use them for our images. For this examination, we have used InceptionV3 pretrained model which has been already trained on Imagenet..The highlights of the pictures from the Flickr 8K dataset are gotten utilizing the Inception v3 model and we have picked this for the better exhibition of the model in object distinguishing proof. Since it is exact and better than VGG16. The PreProcessing will pick up the captions from text files based on the startSeq and endSeq .The Inception is a convolutional neural organization that comprises 14 layers, as this model engineering catches on quickly. These will be prepared by a Dense layer to deliver a 2048 vector portrayal of the photograph and gave to the LSTM layer.

**Defining the Model**

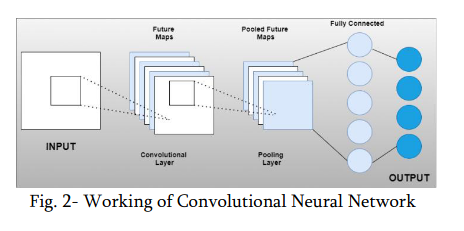
For outlining the shape of our version, we will be the usage of the keras model from functional api. It has three primary steps:

• processing the collection from the textual content

• extracting the characteristic vector from the photograph

• interpreting the output with the aid of concatenating the above layers.

Convolutional Neural Networks (ConvNets or CNNs) are a type of Artificial Neural Network that has shown to be very good at image recognition and classification. Object detection, self-driving cars, image captioning, and other tasks have all made extensive use of them.. The diagram below depicts a basic convnet:

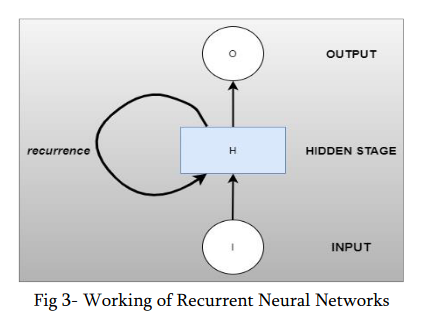


**Figure 2 Convolutional Network(Shukla,2021,p.3)**

Convolution, Non-Linearity (ReLU), Pooling or Sub Sampling and Classification are the four main operations that can be used to explain the entire architecture of a convnet (Fully Connected Layer).Because these operations are the fundamental building blocks of every Convolutional Neural Network, understanding how they work is critical to gaining a thorough understanding of ConvNets.

For Proposed methodology,we will be using different sets of Convolutional networks on the given Flickr8 Dataset.The different networks which are going to be used will be VGG16, AlexNet, ResNet and Custom CNN. The dataset will be trained on different CNNs and the performance of each CNN will be taken into consideration so that the best model to feed to RNN be used. BLEU Score metric will be used for evaluating generated sentences with reference sentences.Based on use cases of different CNNs we will select the model for image captioning and the target to achieve for next dataset will be Flickr30-dataset which will be challenging task to achieve.Also different architectures for training the model or captioning will be used such as CNN-LSTM,Attention architecture etc.

A recurrent neural network (RNN) has the ability to process arbitrary length sequences and is thus used to generate text sequences. We used a Flicker8 dataset for this model. A model trained to predict the caption based on image. By repeatedly calling the model, longer text sequences can be generated. The model demonstrates how a character- based RNN can be used to generate text.It is trained on small batches of text and can produce a longer sequence of text with a logical structure. The Tensorflow library is used in conjunction with other libraries like numpy and os. After that, the dataset is downloaded and read. After that, the data is vectorized, which means the strings are mapped to a numerical representation before training. The model receives a sequence of characters as input, and we train it to predict the output. The text was split into manageable sequences using tf.data, and the data was shuffled and packed into batches.Fig 3 depicts the RNN:



**Figure 3. Recurrent Neural Network**(**Shukla,2021,p.5)**

A Recurrent neural network (RNN) is a sort of sequential network where output from the previous step is fed to the current step.RNN utilizes consecutive information or time- arrangement information. It will provide the output generated from CNN to the current step which will be processed for images. Based on the different architectures used the RNN network will be modified and tested on the huge dataset with different CNNs and models.The trained model will be tested and feeded to RNN and the BLEU Score Metric will be calculated,which is a metric for evaluating generated sentence to a reference sentence.As seen in Figure 3. The output of previous state will be feeded to current step with forward feeding and the captions will be generated based on the images.

**CONCLUSION**

We will create a model to caption the images in our project. Our model will also be expanded by captioning the images based on different architectures used for Image Captioning. The research on dataset of Flick8k will give the basic understanding of how to train and define the model with different architectures in picture. We learned how deep learning techniques work and how to build models using them. The dataset contained 8000 images with 5 captions for each image. In this paper, we will introduce a multi-model Neural Network that consequently figures out how to depict the portrayal of pictures. Also how to train model on different architectures.The initial model will be trained on the architecture of CNN and RNN in light of LSTM units producing a depiction sentence appropriate to the picture. Each expression of the depiction will consequently intended for various items that show up in the information picture when it is created. The CNNmodel will be based on making the inscriptions for the information pictures with VGG16,ALEXNET,ResNET and custom CNN. Based on the models BLEU score metric the model will be selected such that it will be used on higher datasets.This model can be utilized for an assortment of uses, for example, assisting the outwardly disabled individuals with understanding the climate wherein they are in by utilizing text to discourse transformation instrument and naming the pictures that are put on inward efficiently and effectively. In this, we examined the CNN model, RNN models, LSTM models, and eventually, we approved that the model is making subtitles for the information pictures. Further,the model will be trained based on different architectures and the different datasets for model to perform well.

**REFERENCES**

[1] Sujeet Kumar Shukla, Saurabh Dubey, Vineet Mishra “Image Caption Generator Using Neural Networks in Computer Vision” – ECCV 2021 Eds., pp. 15–29, Springer, 2021.

[2] B. Nikitha,Vaibhav Julakanti,G Prasad Acharya, “Image Caption Generator using CNN-LSTM Deep Neural Network”, 2021.

[3] Jitesh Gupta,Mohd Zeeshan,Karan Garg, “Image To Caption Generator”, International Journal for Modern Trends in Science and Technology, 2020.

[4] Marc Tanti, Albert Gatt, Keneth Camillery, “Where to put The Image in Image Caption Generator”, Institute of Linguistic and Language Technology, 2018.

[5] Oriol Vinyals, Alexander Toshev, Sammy Bengio, and Dumitru Erhan, “Show and Tell:A Natural Image Caption Generator,” 2015.