Automated Image Description: Deep Learning-Powered Visual Caption Generation with CNN and LSTM

Sanjana Adagimath
dept. of Electronics and Communication
KLE Technological University
Hubli, Karnataka
sanjanamadagimath@gamil.comn

Gourishankari S Patil
dept. of Electronics and Communication
KLE Technological University
Hubli, Karnataka
gourishankarip@gmail.com

Sarvesh K
dept. of Electronics and Communication
KLE Technological University
Hubli, Karnataka
01FE22BEC468@kletech.ac.in

Ramesh Tabib
dept. of computer science engineering
KLE Technological University
Hubli, Karnataka
ramesh_t@kletech.ac.in

Abstract— In this paper, we propose Deep Learning a novel approach for generating descriptive captions for visual images using a deep learning-based model. In this method it combines Convolutional Neural Networks (CNNs) for image feature extraction and Long Short-Term Memory networks (LSTMs) for sequential data processing to enhance the process of visual image caption generation. The scope for this research spans in multiple domains including healthcare, autonomous vehicles, and entertainment. The motivation behind this is to increase the need for image understanding systems in various applications. To address this, we propose a model which employs a CNN as encoder to extract meaningful visual features and an LSTM as decoder to generate well organized and contextually a applicable captions. We demonstrate the results of the image caption generator using a flickr 8K dataset and to generate accurate captions for the given images.

Keywords— CNN, LSTM, BLEU, VGG16, Image captioning, Deep learning. (key words)

I. Introduction

The ability of the human brain to quickly grasp and accurately describe an image is something that has been difficult to replicate in computers. However, with recent advancements in computer vision and deep learning, it is now possible to train a machine to process and label an image with a highly relevant and accurate caption. Generating a proper sentence to describe the image is still a challenge, but with the right techniques and algorithms, it is possible to build a caption generator that can produce accurate results. The task of image captioning involves identifying the objects in an image and finding the appropriate words to describe them. To form a caption, these words are combined to create a sentence that accurately describes the image. This process requires a combination of computer vision and natural language processing techniques, as it involves both understanding the content of the image and being able to properly describe it in natural language. The model is trained on multiple sentences and images, so that it can learn to generate a wide range of captions for different images with different objects. In this paper, we investigate a method for generating image captions using deep neural networks. In particular, we employ Convolutional Neural Networks (CNN) and Long Short-term Memory (LSTM) to examine the image and produce the description. The objective is to input a picture and output a sentence that accurately describes its contents, with proper grammar.

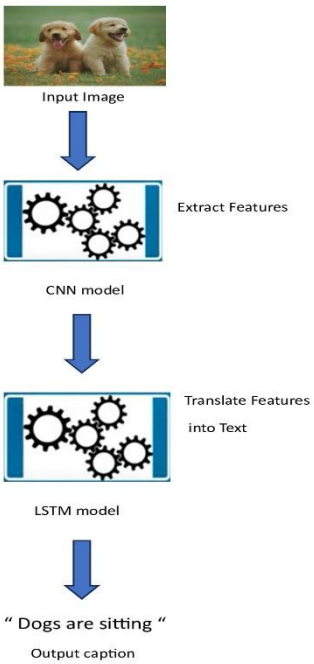

Figure 1: The figure highlights the methods used to generate a caption.

II. LITERATURE SURVEY

[1] Image Caption Generator using Deep Learning :2022

The literature survey in the document discusses the utilization of big data and machine learning for describing images, the use of CNN and RNN for image captioning, and the application of deep learning models for generating captions for input images. It also mentions the use of LSTM and CNN for feature extraction and caption generation, as well as the deployment of the proposed model using the Anaconda framework. Additionally, the survey highlights the importance of diverse and natural captions in image captioning models and the potential applications for visually impaired individuals. The document also references various research papers and conferences related to image caption generation and deep learning techniques.

[2] A parallel-fusion RNN-LSTM architecture for image caption generation

This paper presents a novel parallel-fusion RNN-LSTM architecture. In this paper the strengths of CNN and RNN are combined and the RNN hidden units are divided into several parts of same size and makes them work in parallel. The outputs are merged with corresponding ratios to generate final results with increased efficiency. The proposed model outperforms Google NIC in BLEU and Log Bilinear in Meteor on Flickr8k without the use of additional training data or structures.

[3] Detection and Recognition of Objects in Image Caption Generator System: A Deep Learning Approach -2019

In this paper Regional Object Detector (RODe) is used for object, people and animal detection, recognition and for caption generation. The proposed approach focuses on deep learning to further enhance the existing image caption generator system. CNN is being used for feature extraction and for scene classification. RNN (Recurrent Neural Network) for human and object attributes. The dataset used here is Flicker8k

[4] Captioning Image Using Convolutional Neural Network (CNN) and Long-Short Term -2019

In this paper image caption generation is done using LSTM-based recurrent neural networks and Convolutional Neural Networks (CNN). They have trained their model using Flickr8k dataset. specifically focusing on predicting whether images contain pornographic content. The BLEU evaluation metric is being used.

[5] A new CNN-RNN framework for remote sensing image captioning-2020

This paper focuses on generating captions for Remote sensing (RS) images. They have used a new approach that combines two methods: generating captions and retrieving existing ones. It first employs CNN-RNN framework combined with beam search to generate multiple captions for given input image. Later the most appropriate caption is selected by comparing it to the descriptions of similar images. The dataset used here is RSCID.

[6] Image Captioning Using Inception V3 Transfer Learning Model-2021

In this paper the model is trained using Flickr8k dataset here the combination of Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) units to achieve this, the inception V3 model utilizes different NLP strategies for perceiving and clarifying an image meaning in a natural language. The Inception V3 image caption generator model uses CNN (Coevolutionary Neural networks) for image feature extraction and used RNN to generate meaningful descriptions about the object.

[7] Automatic Caption Generation via Attention Based Deep Neural Network Model-2021

In this paper the model is trained using MSCOCO dataset where in Convolutional Neural Network (CNN) is used for image feature extraction and a Gated Recurrent Unit (GRU) for the text decoder added with a local attention module, the constructive of attention-based models is for producing high-quality image captions. They used BLEU as their evaluation metric.

[8] Generating Image Captions using Deep Learning and Natural Language Processing-2021

In this paper the model is trained using Flickr8k dataset. Combining Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN), they specifically focused on Long Short-Term Memory (LSTM), where in the CNN is used to recognize objects and RNN is used to generate meaningful descriptions of the image, which mainly emphasizes the importance of automating image captioning in various domains.

III.METHODOLOGY

A. Data set:

The Flickr8k dataset is widely available and serves as a benchmark for evaluating image-to-set procedures. There are 8,000 pictures here, and each one is accompanied by five descriptions. The photos in this gallery were culled from various sources on the photo-sharing website Flickr. The captions for these photos give excellent descriptions of the people, places, and things depicted. The dataset is more extensive because it depicts non-famous persons and places in addition to well-known individuals and landmarks. There are a total of 6,000 images spread

between the training, development, and testing phases. Advantages of this dataset for this task include: having several labels for a single image boosts the model's commonality and reduces overfitting. The versatility of the model used for image processing and analysis is bolstered by the large variety of training images used to train the model.

B. Steps to build the image caption generator

-Bring all necessary packages into play.

-Make data cleansing.

-Take the feature vector out.

-Loading the dataset for the model training

-Vocabulary checking

-Build a data generator

-The CNN-LSTM model definition

-Train the image caption generator

-Testing the model of the image caption generator.

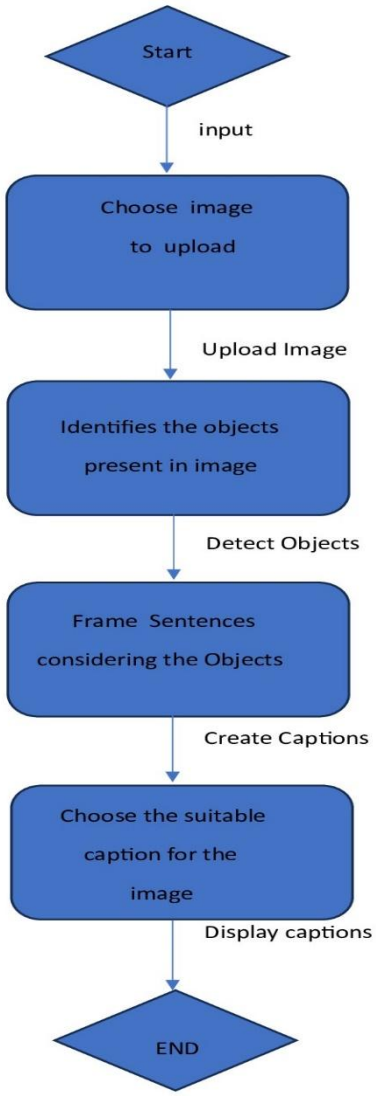

Figure 2: Flow chart

IV. SYSTEM ARCHITECTURE

A. Convolutional Neural Network:

A Convolutional Neural Network is a type of artificial intelligence model that focuses on analyzing visual data. It processes images by dividing them into smaller parts and recognizing the repeating patterns in those parts. This enables the model to identify objects and features in images and make predictions based on that information. It works by applying a set of filters to the input image at various resolutions. These filters are used to extract features from the image, such as edges, shapes, and colors. The CNN then uses these features to build a representation of the image, which is used to identify objects or patterns within the image. Based on its training data, the CNN is thus able to extract objects present in the given image. CNNs allow the extraction of meaningful information from an image which constitutes the first part of the caption generation process. Thus, CNNs are an effective tool for image captioning. The convolution layer applies filters to the image to identify specific patterns and features. The activation layer helps to rectify non-linear relationships by transforming negative values to zero. The pooling layer then simplifies the output by reducing its dimensionality through down-sampling. This results in a more compact representation of the information and helps to reduce overfitting.

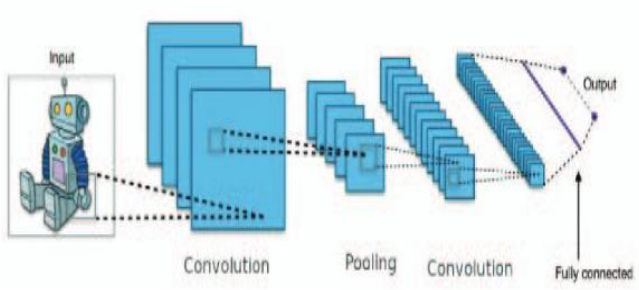

Figure 3: The figure describes how the CNN works.

B. Long Short Term Memory :

An LSTM network is a kind of recurrent neural network which is capable of remembering previous inputs for a certain period of time, which allows it to handle sequential data such as time series, text, and speech. An LSTM network operates using two main components, the memory cell and the gates, to control the flow of information. The memory cell acts as a storage unit that holds information over a prolonged period, while the gates regulate the transfer of information into and out of the memory cell. The gates decide what information should be added to the memory cell, what should be discarded, and what should be passed on to other parts of the network. An LSTM model is designed to maintain a memory of the information it has processed and use hat memory to inform its current processing. In our image caption generator model, we will

combine these two network architectures, which is commonly referred to as a CNN-LSTM model.

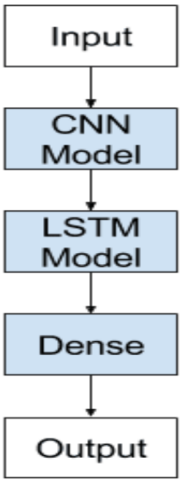

Figure 4: The figure shows the architecture for CNN-LSTM model.

C. VGG16

VGG stands for Visual Geometry Group: it is a standard deep Convolutional Neural Network (CNN) architecture with multiple layers. It is a popular variant of CNN and is widely considered to be one of the most advanced models for computer vision tasks. The VGG16 model is a highly advanced convolutional neural network that was created with the goal of improving upon previous computer vision models. The creators of this model experimented with different architectural designs and ultimately decided to use a deep network with small convolution filters. This resulted in a model with a large number of layers and trainable parameters, and ultimately led to VGG16 becoming one of the best-performing models in the field of computer vision. It is used for classification and identification of images belonging to diverse categories.

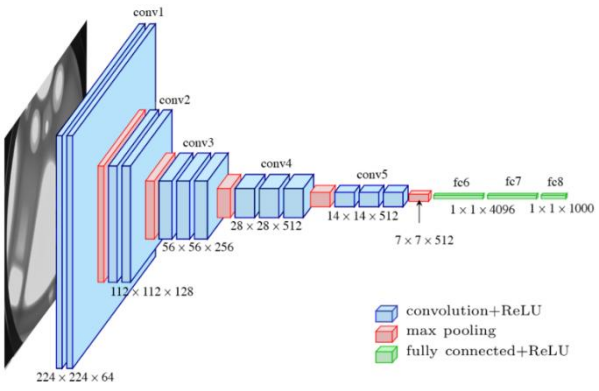

Figure 5: The figure shows how vgg16 works.

D.CNN-LSTM architecture:

The CNN-LSTM architecture is a combination of two different types of neural networks: CNN and LSTM. The CNNs are considered very efficient for performing recognition tasks for components of an image. CNN layers extract characteristics from the input data, while LSTMs generate predictions for sequences of data. Image captioning is related both to computer vision as well as language

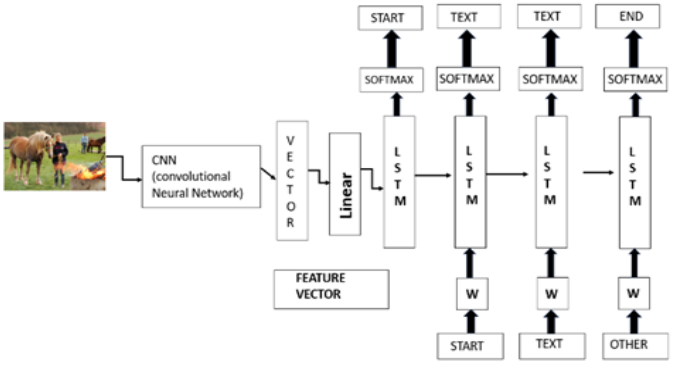

Figure 7: This figure shows how the training data is used.

The CNN-LSTM architecture is particularly useful when the input data has both spatial and temporal structure. Spatial structure refers to the placement of components in a particular space, such as the arrangement of pixels in an image or words in a sentence, paragraph, or document. This technology is also useful in satellite imaging for analyzing deforestation. This architecture is also used when the output is expected to have temporal structure, like the sequence of words in a textual description. In summary, CNN-LSTMs are used when the input data has both spatial and temporal structure, and the output is also expected to have temporal structure.

V. IMPLEMENTATION

The flickr 8k dataset is a collection of 8000 jpeg format images and each image has been given a text descriptions. It has 2 main directories: one is collection of images and other containing text files with detailed descriptions. This file holds the names of the images and their corresponding captions, with each pair being separated by a new line character ("\n"). For every images in the dataset collected has five English captions. The dataset is categorized into 6000 training dataset, 1000 testing dataset and 1000 development datasets images. The "Flickr8k.token" file is undergoing a process of cleaning and formatting as it depend on usability of text descriptions. The dataset has been organised in a sequential pattern in a specific order and tqdm tool is used to track the progress bars in python and to visualise the progress of tasks.

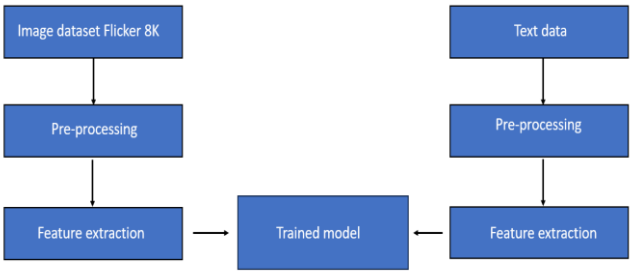

Figure 7: This figure shows how the training data is used.

Our model works like Imports necessary libraries, including TensorFlow, Keras, and others. Mounts Google Drive using Google Colab. Loads the VGG16 model and extracts features from the second-to-last layer. Iterates through a directory of images, preprocesses each image, and extracts features using the VGG16 model. Stores image features in a dictionary. Processes captions by converting to lowercase, removing digits and special characters, adding start and end tags. Tokenizes captions using the Keras Tokenizer. Creates a data generator for training the model by generating batches of image and text data. A model with CNN as encoder (for image features) and LSTM as decoder (for text generation). Compiles the model using categorical cross entropy loss. Checks for missing keys in the training data and adds default features for missing keys. Trains the model using the data generator for a specified number of epochs. It defines functions to convert token indices back to words and to generate captions for given images using the trained model. Evaluates the model using the BLEU score on a test dataset and defines a function to generate and display captions for a given image.

The binary cross-entropy, often employed as a loss function in machine learning, particularly in binary classification problems, measures the difference between predicted and actual probabilities binary outcome.

Binary Cross-Entropy Loss = -(y.log(p)+(1-y).log(1-p))

Where:

Y=ground truth label (0 or 1),

P=predicted probability of input belonging to class1.

Evaluation metric:

BLEU is a method to measure the similarity between a generated sentence and a reference sentence. It is mostly used to evaluate the performance of machine translation systems. The score ranges from 0.0 to 1.0, where 1.0 represents a perfect match and 0.0 represents a complete mismatch. To evaluate the performance of the model, the generated captions are compared with the actual captions. The similarity between these two sets is measured using BLEU score, which is a method to evaluate text similarity. This score is calculated for the entire set of captions, and it provides a summary of how well the generated captions match the expected captions.

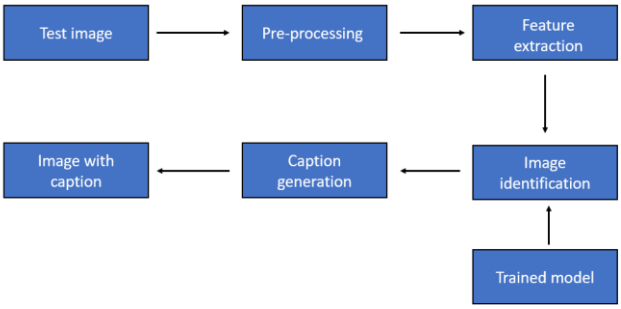

Figure 8: This figure shows how the testing data is used.

VI. RESULTS

Two images are used as input in the proposed model to generate the associated captions. The input images along with respective title are shown in Fig. 9 and 10. The result in terms of generated captions shows accuracy and reliability of the proposed model.

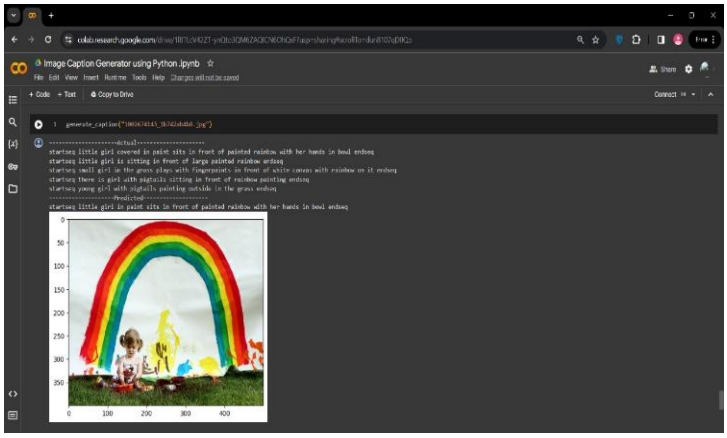

Figure 9: The output given for the generated image.

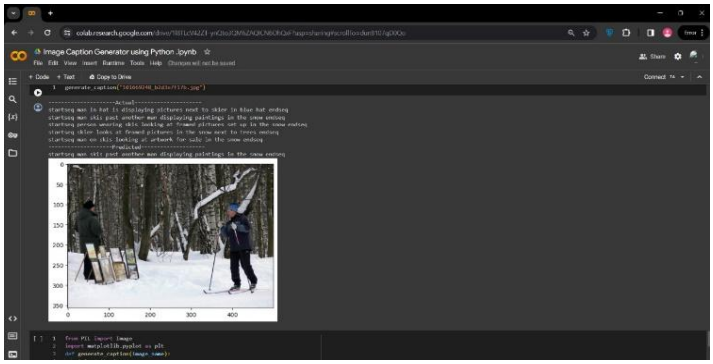

Figure 10: The output given for the generated image.

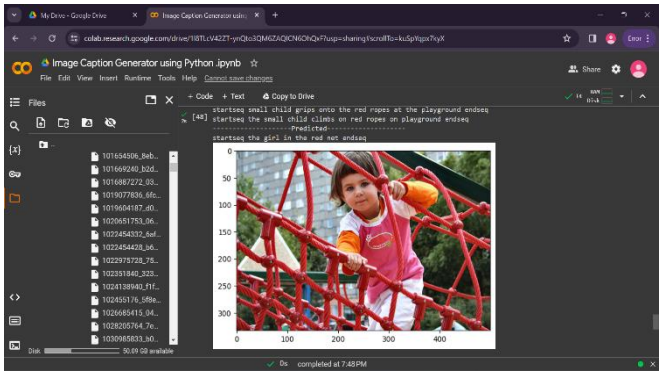

Figure 11: The output given for the generated image.

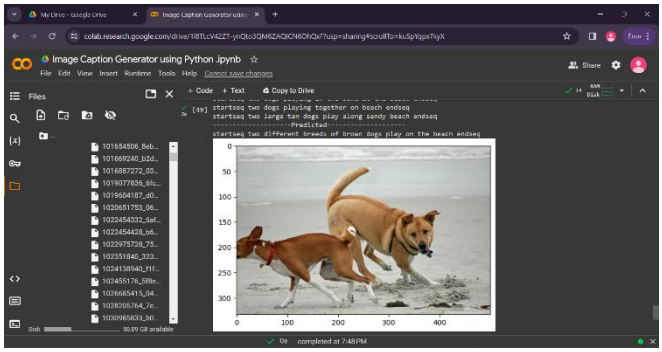

Figure 12: The output given for the generated image.

VII. CONCLUSION

In conclusion, our implementation of a Visual Image Caption Generator utilizing a deep learning model, combining CNN and LSTM, has shown good results. The dataset was relatively small sample size 8000 images, were we could work it on our normal laptop but as compared to business-level model which gets benefit from training a larger number of datasets exceeding 100,000 images.

Our approach, involving a CNN for encoding captions and an LSTM for decoding within a multi-label classification model framework, demonstrated its capability to generate contextually relevant captions. Looking ahead, future developments could include describing captions based on various targets and providing captions in multiple languages. We have our accuracy 69.8%. Expanding the dataset size is crucial for achieving even more accurate results. Therefore, future endeavour will involve testing the model on a larger dataset, aiming to refine the Visual Image Caption Generator for greater precision and broader applicability.

REFERENCES

[1] Generating Image Captions using Deep Learning and Natural Language Processing-2021[8]

[2] Automatic Caption Generation via Attention Based Deep Neural Network Model-2021[9]

[3] Image Captioning Using Inception V3 Transfer Learning Model-2021[6]

[4] Intelligence Embedded Image Caption Generator using LSTM based RNN Model-2021[5]

[6] Context-based Image Caption using Deep Learning-2021[3]

[7] Detection and Recognition of Objects in Image Caption Generator System: A Deep Learning Approach -2019[4]

[8] Seung-Ho Han, Ho-Jin Choi (2020): Domain-Specific Image Caption Generator with Semantic Ontology.

[9] BaiShuang, and Shan An. "A survey on automatic image caption generation." Neuro computing 311 (2018)

[10] C.Szeged, V.Vanhoucke, S.Ioffe, J.Shlens, and Z. Wojna, "Rethinking the inception architecture for computervision," IEEE Conference on Computer Vision and Pattern Recognition, pp. 2818-2826, 2016.

[11]J.Donahue, Y.Jia,O.Vinyals, J.Hoffman, N.Zhang, E. Tzeng, and T.D.Decaf, "A deep convolutional activation feature for generic visual identification," International conference on machine learning, pp. 647-655, 2014.

[12] P. Ajay, B. Nagaraj, R. Arun Kumar, Ruihang Huang, P. Ananthi, "Unsupervised Hyperspectral Microscopic Image Segmentation Using Deep Embedded Clustering Algorithm",2022