IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 11, Issue 12, December 2023

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.379

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 11, Issue 12, December 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1112019 |

Image Caption Generator Using Deep Learning

Hule Sakshi G[1], Wajage Kshitija P[2], Thorat Pratik R[3], Dr. A. A. Khatri[4]

B.E Students, Department of Computer Science and Engineering, Jaihind College of Engineering Kuran, Pune, Maharashtra, India[1,2,3]

HOD, Department of Computer Science and Engineering, Jaihind College of Engineering Kuran, Pune, Maharashtra, India[4]

Abstract: Due to the growing amount of user-generated material on many platforms, including social media, e-commerce websites, and online reviews, image sentiment analysis has attracted a lot of attention. Understanding public opinion, customer happiness, and decision-making processes all depend on the ability to accurately identify sentiment in these texts. This paper suggests a novel method for context-aware sentiment analysis that makes use of a hybrid deep learning model and a variety of feature extraction approaches. Utilizing several feature extraction approaches, such as bag-of-words, word embeddings, and syntactic dependency parsing, to first extract the contextual information from the text. These methods make it possible to represent the text in a structured way that considers both the semantic and syntactic elements. Our model can comprehend the sentiment more effectively by including these features. Our method's foundation is the deep learning model, which combines the advantages of long short-term memory (LSTM) networks and convolutional neural networks (CNNs). While the LSTM component catches long-term relationships and preserves contextual data, the CNN component captures local dependencies and learns high-level characteristics. Our approach successfully captures both local and global context by combining these two elements, which enhances sentiment analysis performance. Prior to creating the visual feature vector for the generation of captions, the execution first chooses the context. To provide the image description for each unique photo, the EfficientNetB7 model is used. The two methods used in the classification of sentiment labels are the attention-based LSTM and the greedy Gated Recurrent Unit (GRU) method.

KEYWORDS:- Image Caption, Convolutional Nural Network, Long Short Term Memory, Natural Language Processing

I. INTRODUCTION

It is not difficult for us to put ourselves in another person's shoes when a system is observing them, or even to feel, at least in part, the feelings that they appear to be feeling. Our unique ability to infer the emotions of those around us is extensively used during our daily lives. Our ability to empathize with others makes it possible for us to interact with them in social situations in a more helpful, sensitive, empathic, affectionate, and pleasant manner. In a broader sense, this capacity allows us to comprehend the motives and goals that underlie other people's behaviour and to predict how they will react to situations.

The programme combines CNN and RNN, two key architectures that define properties, connections, and objects in images and translate them into English. CNN is an extractor that takes features out of the image that is provided. The CNN output will be passed into the RNN-LSTM, which will then describe and provide a caption. Convolutional neural networks, or CNNs, process data with input shapes resembling two dimensional matrices. The input layer, convo layer, pooling layer, fully-connected layers, softmax, and output layers are only a few of the numerous layers in the CNN model. An image serves as CNN's input layer. The presentation of image data takes the form of a 3D matrix. Convolutional layer, also referred to as a feature extractor, performs convolutional operations and computes dot products. All negative values are converted to zero by the ReLU sublayer of the Convo layer. After the convolution layer has been applied, the volume of the picture is lowered in the pooling layer.

Fully-Connected layers are a type of connection layer that uses neurons, biases, and weights to connect one neuron in one layer to another neuron in another layer. Use of the Softmax layer allows for the multi-classification of objects utilising formulas. The encoded result is sent to the LSTM model in the output layer, which is the final layer of the CNN model.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 11, Issue 12, December 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1112019 |

II. LITERATURE REVIEW

Convolutional Neural Network models, as per [1] have played a critical impact in this picture. Here, we endeavor to exhibit and feature a few procedures for picture highlight extraction, as well as how they will be utilized to subtitle age. In the space of information science, there has been a ton of examination towards further developing picture subtitle creating models. Normal language handling is critical in delivering a portrayal that is exact and has meaning. The leaned toward network in depiction development is the repetitive brain organization (RNN), which is generally used for grouping age. Scientists have invested a ton of energy and work to make monstrous information bases. The MSCOCO dataset, which is provided by Microsoft, is one of the most notable datasets. The Flickr 8K, Flickr 30K, PASCAL, and a couple of more are extra notable and benchmark datasets. Subtitling pictures, as per [2] is the demonstration of making clear data about visual items, picture metadata, or things that exist in an image. The material inside the photos is extremely valuable according to the point of view of PC vision. They help a machine's cognizance and execution. Picture subtitling has different purposes, including altering programming ideas, menial helpers, picture ordering, openness for outwardly impeded individuals, interpersonal interaction, and an assortment of other normal language handling applications. The course of picture explanation has been achieved at a more significant level and has added to various regions through various techniques for profound learning arrangement. Individuals have concocted different imaginative ways of moving toward this application utilizing profound learning. It has been shown that profound learning models are fit for accomplishing ideal results in the space of subtitle producing issues in view of these discoveries. It's basic to grasp not exactly what the things in the picture are, yet additionally the way that they connect with each other, to deliver great picture depictions.

Encoder-decoder models are presently respected one of the most exceptional picture inscribing techniques.

A great deal of data is put away in a thing. Immense measures of picture information might be created every day on long range interpersonal communication destinations, including galactic items, yet this is an up with the speedy thing. Commenting on pictures of individuals takes more time, and the possibilities committing an error are higher. Profound learning models are used to develop such pictures accurately, eliminating the requirement for human changes [3]. By disposing of the necessity for human investment, this would considerably diminish human disappointment and exertion. The improvement of picture explanations has various true advantages, going from helping the slow-witted to helping the computerized, financially savvy stamping of pictures shared web-based consistently, rules for handling programming, valuable for brilliant gadgets, picture encoding, outwardly impaired individuals, person to person communication destinations, and an assortment of other normal writing. Bricklayer and Charniak use visual closeness to get an assortment of subtitled pictures for a question picture [4] to moderate the impacts of loud visual evaluations in procedures that depend on picture recovery for picture inscribing. They then gauge a word likelihood thickness molded on the inquiry picture utilizing the inscriptions of the recovered pictures. The term likelihood thickness is utilized to survey existing subtitles to pick the one with the most elevated score as the question's inscription. This procedure has certainly expected that there is consistently an expression that is pertinent to a question picture. Truly, this supposition that is rarely precise. Rather than straightforwardly using returned sentences as depictions of question pictures, recuperated sentences are utilized to build another portrayal for an inquiry picture in a different line of recovery based research.

Li et al. use visual models to extricate semantic data from pictures, including articles, qualities, and spatial associations [5]. Then, for encoding acknowledgment results, they build a trio of the kind adj1, obj1, prep, adj2, obj2. To give a depiction to the trio, web-scale n-gram information is utilized to lead express determination, which might give recurrence counts of potential n-gram successions. This considers the assortment of potential expressions that might make up the trio. Following that, state combination is utilized to use dynamic programming to find the most reasonable assortment of expressions to act as the question picture's depiction.

III. METHODOLOGY

DATASET:

- For the application of image caption generator, we used the dataset named, Flickr 8k dataset.
- This dataset contains a wide range of images that has many different types of situations and scenes.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 11, Issue 12, December 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1112019 |

- Flickr 8k dataset has 8000 images and every image has 5captions.
- We divided the entire dataset of 8000 images as 6000, 1000and 1000 as training, validation and testing sets respectively
- Every image has different dimensions.

A. Dataset Collection:

Data is collected from a variety of sources and prepared for data sets. And this data is used for descriptive analysis.

B. Preprocessing step:

This step is a very important step in machine learning. Preprocessing consists of inserting the missing values, the appropriate data range, and extracting the functionality.

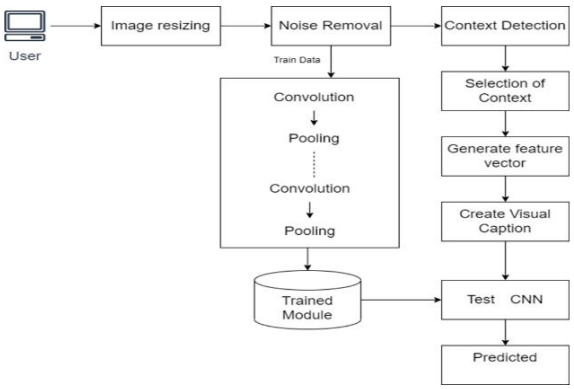

Fig. 1. System Architecture

C. Feature Selection:

Feature extraction should simplify the amount of data involved to represent a large data set. The soil and crop characteristics extracted from the pre-treatment phase constitute the final set of training.

D. Data Prediction:

In Advance to this step there need to split the data into train dataset and test dataset. By applying the CNN algorithm the data is trained with available input and output data. Then the new data is predicted by machine learning modules

IV. PSUDO CODE

System Description:

- Input=product, gesture
- Output=detected gesture
- Let S be the | System as the final set
- S = {U,Fr,G,A,F}

- Let U be the set of Users where,

– S = {U}

– U ={U1, U2, U3, U4.....| U }

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 11, Issue 12, December 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1112019 |

– where, U1=user1.

– U2=user2.

• Let Fr be the set of Frame where,

– S = {Fr}

– Fr = {Fr1, Fr2, Fr3.....| Fr }

– where, Fr1=Frame1

– Fr2=Frame2

• Let G be the set of Gesture where,

– S = {G}

– G = {G1, G2, G3.....| G}

• Let A be the set of Algorithm where,

– S = {A}

– A = {A1, A2, A3.....| A}

• Identify the functions as ?F

• S = {F}

• F = {F1 (), F2(), F3(), F4(), F5(), F6(), F7(),F8(),F9()}

– F1 (S) = Grab Image

– F2 (S) = define gesture

– F3 (S) = Grey Scale

– F4 (S) = Threshold

– F5 (S) = Bluring

– F6 (S) =Image substraction

– F7 (S) =Gesture Detection

– F8 (S) =Add Product

– F9 (S) =Manage Product

• Success condition-cloth fitted to customer.

• Failure condition-Cloth is not fitted.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 11, Issue 12, December 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1112019 |

V. PROPOSED SYSTEM

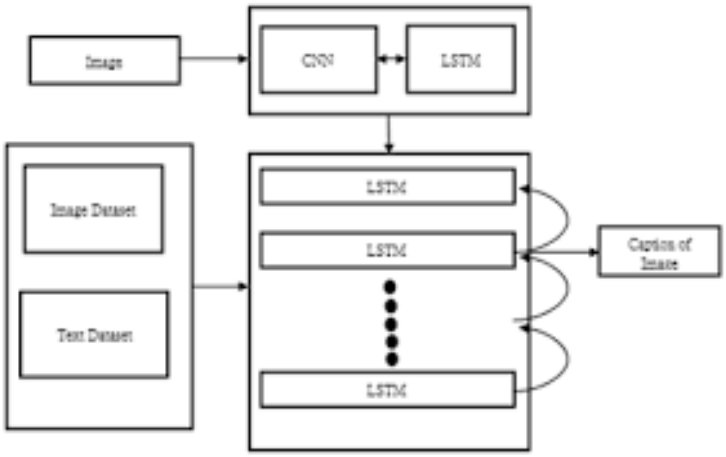

Fig. 2. Propose System

The proposed model of Image Caption Generator is as described in this proposed system, In this model, input image is given & then a convolutional neural network is used to create a dense feature vector. This dense vector, also called an embedding, this vector can be used as input into other algorithms, and its generates suitable caption for given image as output. For an image caption generator, this embedding becomes a representation of the image and used as the initial state of the LSTM for generating meaningful captions, for the image.

VI. RESULTS AND DISCUSSIONS

In a project, image captioning refers to generating textual descriptions or captions for images using machine learning and NLP techniques. It involves analyzing the visual features of an image and generating a coherent and relevant caption that describes the image's content.

Output: Cats sitting on the ground with a pair of shoes

VII. CONCLUSION

Our model based on multi label classification using fast Text and CNN, is useful in detecting and extracting objects from image and generate caption according to the provided datasets. We have presented multiple approaches for Image caption Generator like (Convolution neural network). The CNN model was built on the idea

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 11, Issue 12, December 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1112019 |

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 and will be validating that the model is generating captions for the input pictures.

VIII.FUTURE ENHANCEMENT

Image captioning can be used to improve assistive technology and aid visually impaired to comprehend their environment. The captions generated as output can be read aloud to the users and help them to interact with it better.

REFERENCES

1. S. Yan, F. wu, J. Smith and W. Lu , 2019 “Image Captioning via a Hierarchical Attention Mechanism and Policy Gradient Optimization
2. R. Staniute and D. Sesok , 2019 “A Systematic Literature Review on Image Captioning
3. M. Z. Hossain, f. Sohel, m. F. Shiratuddin and h. Laga , 2018 A Comprehensive Survey of Deep Learning for Image Captioning.
4. D. S. Whitehead, L. Huang, H. and S.-F. Chang“Entity aware Image Caption Generation,” in Empirical Methods in Natural Language Processing, Brussels, 2018.
5. G. Ding, M. Chen, S. Zhao, H. Chen, J. Han and Q. Liu, ”Neural Image Caption Generation with Weighted Training and Reference”, Cognitive Computation, 08 August 2018.
6. J. Chen, W. Dong and M. Li,“”Image Caption Generator Based On Deep Neural Networks”, March 2018.
7. S. Bai and S. An, “”A Survey on Automatic Image Caption Generation”, Neuro computing, 13 April 2018

INNO SPACE
SJIF Scientific Journal Impact Factor

Impact Factor: 8.379

doi cross ref

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details