IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Image Caption Generator Using Deep Learning

[1]Prof.S. Sankareswari, [2]Miss.Bibi Zainab Dongarkar, [3]Miss.Heena Dongarkar, [4]Miss.Simran Sarang, [5]Miss.Madhura Valke

[1]Assistant Professor, [2]Student, [3]Student, [4]Student, [5]Student

[1]Department of Information Technology,

[1]Finolex Academy of Management and Technology, Ratnagiri, India

Abstract: In order to automatically create evocative descriptions for photos, the Image Caption Generator Project introduces a novel blend of computer vision and natural language processing approaches. Convolutional Neural Networks (CNNs) are used by the system to process raw photos while utilizing cutting-edge deep learning models to recognize complicated patterns and objects. This visual comprehension is seamlessly combined with cutting-edge Natural Language Processing (NLP) algorithms, using attention processes and Sequence-to-Sequence models to produce captions that are both linguistically and contextually coherent. The project places a strong emphasis on the user experience by giving users a simple interface via which they can upload photographs and instantly receive pertinent captions. The reliability and correctness of generated captions are guaranteed by stringent evaluation measures like BLEU and METEOR. The system must be trained on a variety of datasets to ensure ethical considerations, minimize biases, and promote inclusive outcomes. Potential applications of the project include search engine content metadata enrichment, accessibility tools for the blind, and boosting user engagement on social media platforms.

Index Terms – Image, Caption, CNN, RNN, LSTM

I. INTRODUCTION

Deep Learning-based image caption generators are advanced AI systems that produce informative captions for photos by fusing computer vision and natural language processing methods. This technique bridges the gap between visual content and human-readable descriptions by utilizing deep learning models, typically convolutional neural networks (CNNs) for image feature extraction and recurrent neural networks (RNNs) or transformer designs for text generation. An image is initially put into the model to start the process. By removing prominent characteristics from the image, CNN is able to gather crucial visual data. These attributes are subsequently put into the RNN or transformer, which sequentially processes them to create a coherent phrase that accurately and appropriately describes the image. These models are trained using enormous sets of paired images and their captions, which helps the machine understand the relationships between language descriptions and visual patterns. This technology serves applications across a range of industries, including search engine content indexing, accessibility solutions for the blind, and enhancing user interfaces on social media and e-commerce platforms. These picture caption generators have become essential in the field of computer vision and AI-driven content creation due to the constant enhancements made to deep learning algorithms and the accessibility of large datasets.

II. OBJECTIVES

1. Automatic Caption Generation
   Develop a deep learning model capable of automatically generating descriptive captions for a wide range of images.

2. Accuracy and Relevance
   Ensure that the generated captions are accurate and relevant to the content of the images. The captions should effectively describe the objects, scenes, and actions present in the images.

3. Natural Language Fluency
   Generate captions that are fluent and coherent in natural language. The captions should be grammatically correct and linguistically appropriate.

4. Handling Ambiguity
   Address the challenge of ambiguous images by generating contextually appropriate captions. The model should be able to handle images with multiple objects, complex scenes, and diverse contexts.

5. Adaptability
   Design the model to be adaptable to various types of images, including different categories and styles. The generator should perform well on both structured and unstructured image data.

6. Multimodal Understanding
   Explore techniques that enable the model to understand not only the visual content but also the relationships between objects and scenes in the images. Multimodal approaches can incorporate both visual and semantic information for better captioning results.

7. Attention Mechanisms
   Implement attention mechanisms to allow the model to focus on relevant parts of the image when generating captions. Attention mechanisms improve the quality of captions by aligning them with specific regions of interest in the images.

8. Evaluation Metrics
   Define appropriate evaluation metrics such as BLEU (Bilingual Evaluation Understudy), METEOR (Metric for Evaluation of Translation with Explicit ORdering), ROUGE (Recall-Oriented Understudy for Gisting Evaluation), and CIDEr (Consensus-based Image Description Evaluation) to quantitatively assess the quality of generated captions.

9. User Experience
   Consider the end-user experience by generating captions that are not only accurate but also engaging and informative. Captions should enhance the overall user experience when interacting with images.

III. LITERATURE SURVEY

In [1] the RESNET-LSTM model is used to generate captions for each of the given images. RESNET is the architecture of Convolution layer. This RESNET architecture is used for extracting the image features and this image features are given as input to Long Short Term Memory units and captions are generated. This image captioning deep learning model is much useful for analyzing the large amounts of unstructured and unlabeled data to find the patterns in those images for guiding the building the software to guide blind and deaf people.

In [2] The CNN and LSTM work together in proper synchronization, they were able to find the relation between objects and images. BLEU (Bilingual Evaluation Under study) scores are used in text translation for evaluating translated text against one or more reference translation. Over the years several other neural network technologies have been used to create image caption generators, e.g. VGG16 model instead of the Xception model, or the GRU model instead of the LSTM model. Furthermore, BLEU score can be used to draw comparisons between these models to see which one provides maximum accuracy.

In [3] paper discuss the challenging task of automatically creating descriptions or captions for images using natural language sentences. The paper highlights the need for computer vision methods to understand the content of the image, and emphasizes the impressive aspect of using single end-to-end model to predict captions instead of relying on complex data preparation or a pipeline of models. The paper also mentions the effectiveness of Recurrent Neural Networks (RNNs) for sequential data modeling.

In [4] paper discusses the task of automatically describing the content of image using Natural language. It highlights potential impact of this task, such as helping visually impaired individuals understand images on the web. The paper also introduces a modified method for generating image captions, which produces comparable results to state-of-the-art methods but requires less training memory.

In [5] paper discusses the development of an image caption generator using deep learning techniques. Aim to build an optimal system that can generate accurate and grammatically correct caption for images. The VGG-16 model used for object recognition and to extract information from images using Convolutional Neural Networks. The complete system consists of three models: feature extraction model, encoder model and decoder model. The paper concludes the accuracy of different feature extraction and encoder models and their influence on sentence generation.

IV. THE EXISTING PROCESS

There are numerous crucial steps involved in creating an image caption generator using deep learning. First, an extensive collection of pictures with associated captions is gathered. The texts are tokenized and turned into numerical indices, while the photos are preprocessed by scaling and normalizing. An encoder and a decoder are built as part of a deep learning architecture. The encoder extracts picture features, frequently using a convolutional neural network that has already been trained. The decoder creates captions word per word, generally utilizing transformer structures or recurrent neural networks. When creating captions, attention techniques may be used to concentrate on pertinent image parts.

By minimizing a loss function that estimates the difference between generated captions and ground truth captions, the model develops the ability to map images to captions throughout the training phase. Using optimization methods like stochastic gradient descent, the model parameters are adjusted iteratively in this process. The effectiveness of generated captions is evaluated using metrics like BLEU, METEOR, ROUGE, and CIDEr. Performance can be improved through fine-tuning and transfer learning, while taking into account moral concerns about biases in generated text. The trained model then becomes a potent tool for picture understanding and description by being utilized to provide captions for fresh, undiscovered photos.

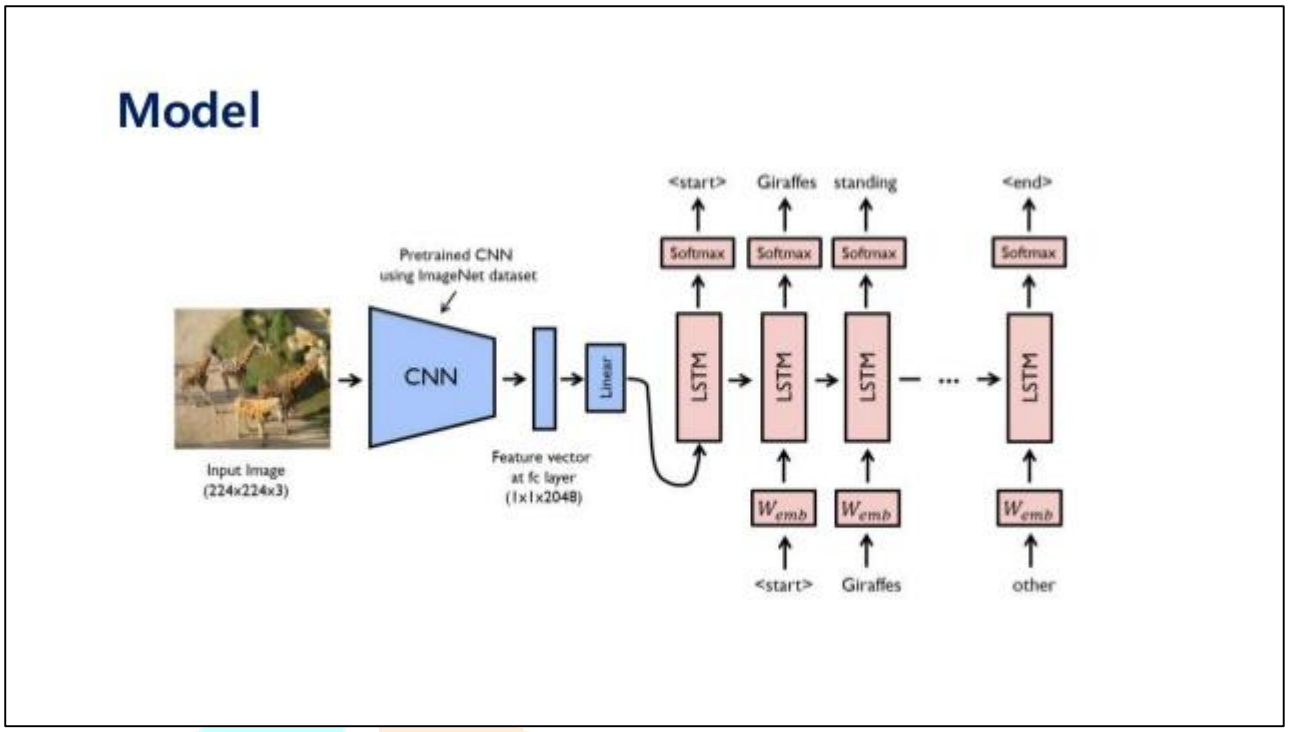

Fig: Image Caption Generator

V. PROPOSED SYSTEM

A. Task

The goal is to create a system that can accept an image input in the form of a dimensional array and produce an output that is a syntactically and grammatically accurate sentence that describes the image.

B. Corpus

We have used the Flickr 8K dataset as the corpus. The dataset consists of 8000 images and for every image, there are 5 captions. The 5 captions for a single image helps in understanding all the various possible scenarios. The dataset has a predefined training dataset Flickr_8k.trainImages.txt (6,000images), development dataset Flickr_8k.devImages.txt (1,000 images), and test dataset Flickr_8k.testImages.txt (1,000 images). The Images are opted from six varied Flickr groups and do not contain any well-known personality or places. However, they are manually selected to show a variety of scenes.

B.Preprocessing

The photos and their related captions are cleaned and treated independently as part of the data pretreatment process. The Xception implementation of the Keras API, which runs on top of TensorFlow, is used for image preprocessing. ImageNet is used to pre-train Xception. This allowed us to use transfer learning to train the images more quickly. The tokenizer class in Keras is used to clean up the descriptions; this vectorizes the text corpus and stores the cleaned data in a different dictionary. Then, a distinct index value is assigned to each word in the lexicon.

D.Model

A hierarchy of levels in an artificial neural network made for deep learning is used to carry out the machine learning process. The model is based on deep networks, where the flow of information begins at the initial level, where the model learns something basic and then passes its output to layer two of the network while combining its input into something slightly more complicated and passing it on to layer three. As each level in the network builds on the information it acquired from the ascending level, this process continues.

Convolutional Neural Networks (CNN)

A Convolutional Neural Network (CNN) is a specialized type of artificial neural network designed for processing and analyzing visual data, particularly images and videos. CNNs excel at recognizing patterns and features in grid-like structures. They employ convolutional layers to apply filters or kernels to input images, capturing local patterns like edges and textures. Activation functions like ReLU introduce non-linearity, and pooling layers reduce the spatial dimensions of the output, focusing on essential information. Fully connected layers at the end of the network integrate high-level features for classification or regression tasks. CNNs revolutionized computer vision by enabling machines to automatically learn and understand intricate visual hierarchies, making them crucial in applications like image recognition, object detection, and image generation. Moreover, CNNs have been influential in other domains like natural language processing and biomedical image analysis. Their ability to automatically learn features from raw data has made them a fundamental tool in modern machine learning, leading to significant advancements in various fields.

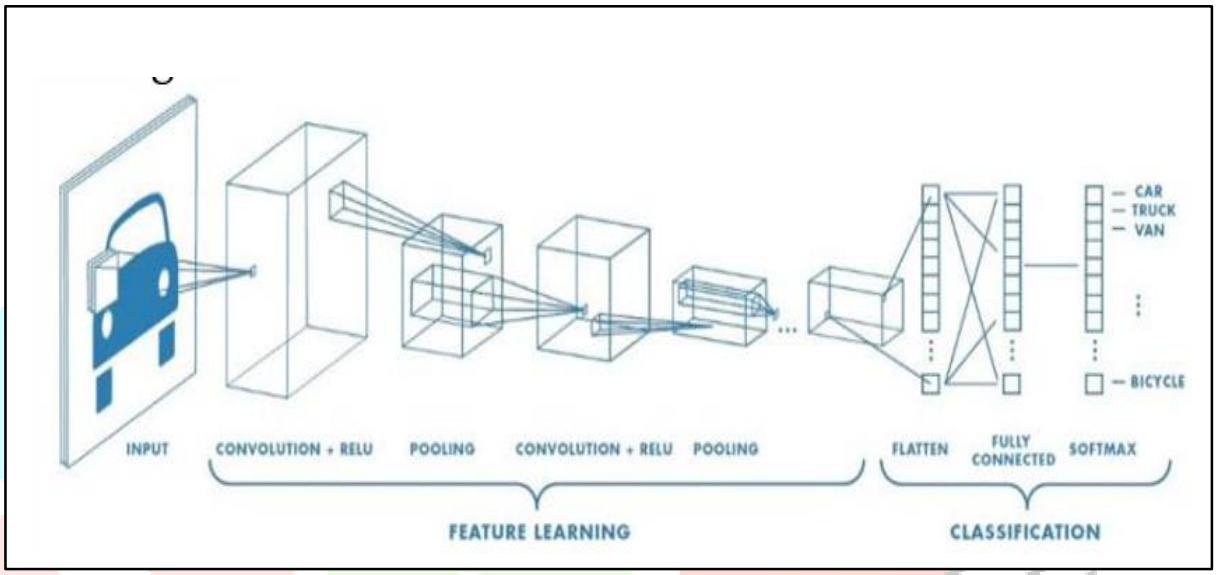

Fig: Architecture of Convolutional Neural Network

Long short-term Memory (LSTM)

Long Short-Term Memory (LSTM) is a type of recurrent neural network (RNN) architecture specifically designed to address the vanishing gradient problem, which hampers the ability of traditional RNNs to capture and learn long-term dependencies in sequential data. LSTMs were introduced to overcome this limitation, making them highly effective for tasks involving time-series prediction, natural language processing, and speech recognition.

LSTMs maintain a cell state—a memory unit that can store and retrieve information over long periods. They have three main components: input gates, forget gates, and output gates. These gates regulate the flow of information into, out of, and within the cell state, allowing LSTMs to selectively remember or forget information. During training, LSTMs learn to update these gates based on the input data, making them capable of capturing patterns and dependencies in sequences of data with varying time lags.

The ability to retain information over extended periods makes LSTMs especially useful in tasks where context from earlier inputs is crucial for understanding the meaning of the current input. For instance, in language modeling, LSTMs can capture the nuances of language, such as sentence structure and meaning, by maintaining relevant context in the cell state. This makes LSTMs a fundamental tool in the field of deep learning, enabling the development of sophisticated models for tasks requiring sequential data analysis.

Architecture

We utilize a CNN + LSTM to take an image as input and output a caption. An "encoder" RNN maps the source sentence (which is of variable length) and transforms it into a fixed-length vector representation, which in turn is used as the initial hidden state of a "decoder" RNN which ultimately generates the final meaningful sentence as a prediction.

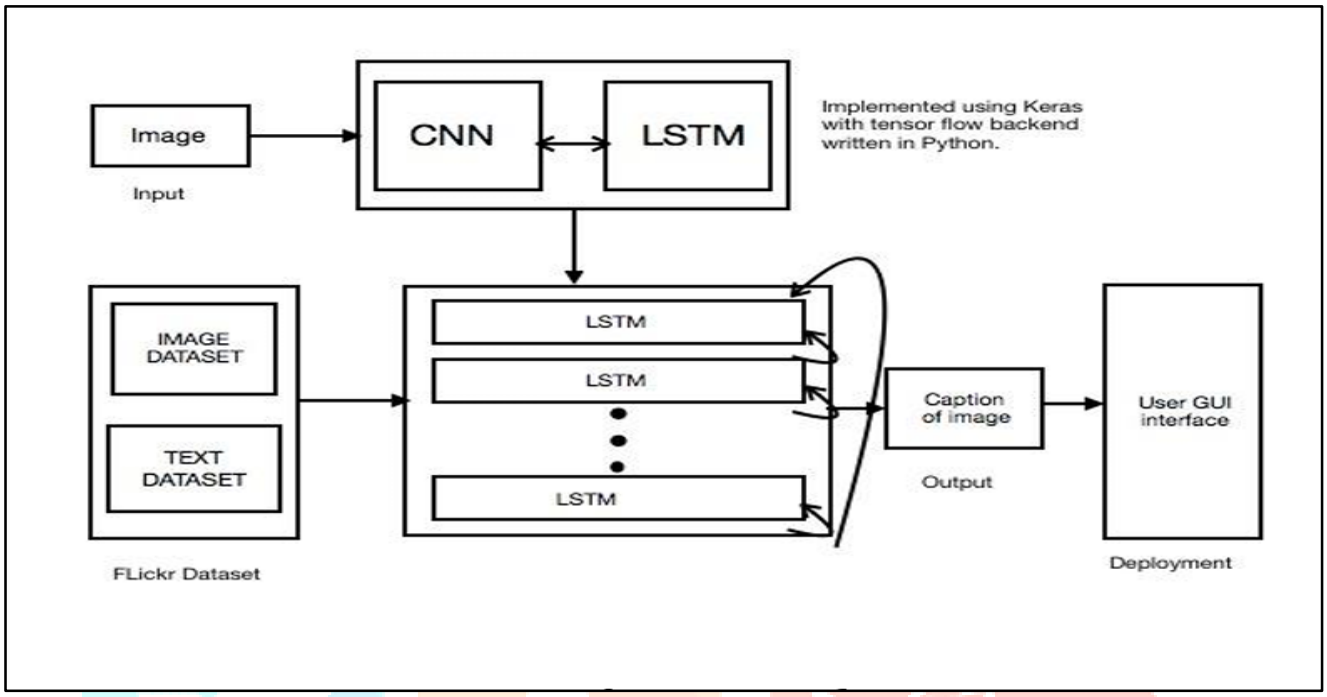

Fig: System Architecture of Image Caption Generator

However, we are going to replace this RNN with a deep CNN - since it can produce a rich representation of the input image by embedding it to a fixed-length vector - 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.

VI. TECHNOLOGY USED

A. Python

Python is a versatile, high-level programming language known for its simplicity and readability. Created by Guido van Rossum and first released in 1991, Python has gained widespread popularity in the software development community. Its ease of use and extensive libraries make it an ideal choice for beginners and professionals alike. Python supports multiple programming paradigms, including procedural, object-oriented, and functional programming, allowing developers to write clear and concise code for various applications. Python's rich standard library provides modules and packages for tasks ranging from web development and data analysis to artificial intelligence and scientific computing. Its simplicity encourages developers to focus on problem-solving rather than dealing with complex syntax, enhancing productivity and speeding up the development process.

B. Jupyter Notebook

Jupyter Notebook is an open-source web application that allows interactive computing and data visualization. It supports various programming languages, with Python being the most popular. Jupyter Notebooks enable users to create and share documents containing live code, equations, visualizations, and narrative text. This interactive environment fosters collaborative and exploratory data analysis, machine learning, and scientific research. Users can run code in a step-by-step manner, making it an invaluable tool

for learning, experimentation, and communication of data-driven insights. Its flexibility and ease of use have made Jupyter Notebook a fundamental tool for data scientists, researchers, and educators.

C. Google Colab

Google Colaboratory, often referred to as Colab, is a cloud-based platform provided by Google that allows users to write and execute Python code in a web-based interactive environment. It offers free access to GPU (Graphics Processing Unit) resources, enabling faster computation for tasks like machine learning and data analysis. Colab integrates with Google Drive, allowing users to create, share, and comment on Jupyter Notebooks directly within their Google Drive accounts. It provides pre-installed libraries for popular data science frameworks, making it a convenient choice for collaborative and resource-intensive projects. With its seamless integration with other Google services, Colab has become a popular choice for researchers, data scientists, and students for collaborative coding and experimentation.

D. Python Libraries

1) Pandas
2) Numpy
3) Matplotlib
4) Keras
5) Re
6) Nltk
7) String
8) Json
9) Pickle
10) TensorFlow

VII. CONCLUSION

Deep learning-based image caption generators mark a substantial development in computer vision and natural language processing. With the ability to reliably and coherently describe complex visual content, these systems demonstrate the synergistic potential of neural networks. These models close the semantic gap between images and textual descriptions by combining recurrent neural networks with attention mechanisms for language creation and convolutional neural networks for extracting image features. Their uses include helping those who are blind and enhancing material indexing and retrieval. Despite the significant progress that has been made, more advanced and inclusive image captioning solutions are still in the works. Current research is focused on improving contextual awareness, correcting biases, and assuring ethical considerations.

VIII. REFERENCES

1. HaoranWang , Yue Zhang, and Xiaosheng Yu, “An Overview of Image Caption Generation Methods”, (CIN-2020)
2. B.Krishnakumar, K.Kousalya, S.Gokul, R.Karthikeyan, and D.Kaviyarasu, “IMAGE CAPTION GENERATOR USING DEEP LEARNING”, (international Journal of Advanced Science and Technology- 2020)
3. MD. Zakir Hossain, Ferdous Sohel, Mohd Fairuz Shiratuddin, and Hamid Laga, “A Comprehensive Survey of Deep Learning for Image Captioning” ,(ACM-2019)
4. Rehab Alahmadi, Chung Hyuk Park, and James Hahn, “Sequence-tosequence image caption generator”, (ICMV-2018)
5. Oriol Vinyals, Alexander Toshev, SamyBengio, and Dumitru Erhan, “Show and Tell: A Neural Image Caption Generator”,(CVPR 1, 2- 2015).
6. Image2Text: A Multimodal Caption Generator by Chang Liu.
7. K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, et al., "Show, attend and tell: Neural image caption generation with visual attention", Proceedings of the International Conference on Machine Learning (ICML), 2015.
8. J. Redmon, S. Divvala, Girshick and A. Farhadi, "You only look once: Unified real-time object detection", Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016
9. D. Bahdanau, K. Cho, and Y. Bengio. “Neural machine translation by jointly learning to align and translate.arXiv:1409.0473”, 2014.
10. Saad Albawi, Tareq Abed Mohammed, and Saad Al-Zawi, “Understanding of a convolutional neural network”, IEEE - 2017