Image Caption Generator

Ayush Sharma
Department of CSE.
KIET Group of Institutions.
Ghaziabad, India
as3400588@gmail.com

Deepesh Singh
Department of CSE.
KIET Group of Institutions.
Ghaziabad, India
deepeshsingh744@gmail.com

Ayush Singh
Department of CSE.
KIET Group of Institutions.
Ghaziabad, India
ayushsinghsln1111@gmail.com

Rahul Kumar Sharma
Department Of CSE
KIET Group of Institutions.
Ghaziabad, India
rahulpccs1988@gmail.com

***Abstract*—One of the most important tools in the modern world is image captioning. Additionally, built-in pro-grams that use deep neural network models to make and produce captions for specific pictures. picture captioning is the process of creating a description for a picture. It necessitates identifying the key elements in an image, their characteristics, and the connections between them. It produces sentences with proper syntactic and semantic structure.In this paper, Our research presents a cutting-edge method utilizing deep learning, machine translation, and computer vision to produce captions that describe images. The primary aim of this investigation is to recognize different elements within an image. Image captioning exemplifies this process effectively. The primary objective of image captioning is to formulate a descriptive statement for a given image. In this research, we introduce a machine learning approach which is working with machine translation and computer vision to produce captions and describe pictures. The objective of this study is to identify various things present in an image, identify the connections among those objects, and produce pictures. In order to demonstrate the proposed experiment, we will utilize a machine learning technique known as Transfer Learning. This approach will be implemented using the machine learning model, together with the dataset and utilizing the Python-language. Index Terms- computer vision, automated generated captions, CNN, machine learning, LSTM**

***Index Terms*—automated captions, computer vision, CNN, machine learning, LSTM**

I. Introduction

The usage of natural language processing (NLP) to come across objects [1] and describe them is an age-vintage problem in synthetic intelligence. This became taken into consideration as a not possible assignment. The development of deep learning techniques, the supply of full-size datasets, and the electricity of computation permit models to be constructed that generate captions for images. photo caption technology is a venture that entails picturestandards for computation and language comprehension to recognize and understand the description of the images and also describe them in a natural like language English language or some other language. The main aim of the project is to craft fitting descriptions, for images. These descriptions aim to encapsulate the details depicted in the images. The current techniques depends on networks (CNNs) and recurrent neural networks (RNNs) or their different variations to create relevant captions of the images.The approach involves utilizing an encoder decoder mechanism, within these networks to accomplish this task. A convolutional neural network processes the image to generate a series of feature vectors while recurrent neural networks serve as decoders and produce descriptions.We used the deep neural networks and machine learning techniques to create a good and accurate model. We used the Flickr 8k dataset containing approximately 8k sample images with five captions per image. There are two phases involved in this process: the initial stage involves extracting features from images through the utilization of a convolutional neural network (CNN), while the subsequent stage involves generating natural language sentences based on the images using a recurrent neural network (RNN) [3]. During the first stage, our approach differs from merely detecting the objects present in the image. Instead, we employ an alternative method to extract features that provide us with intricate details regarding even the most minute distinctions between two similar pictures. To accomplish this, we employ VGG-16 (Visual Geometry Group), a convolutional model consisting of 16 layers that is specifically designed for object recognition purposes.. Deep learning has garnered interest due, to its proficiency in a type of learning that holds promise for real-world use. [1] Having the capability to derive insights from disorganized data is a significant advantage for individuals interested in putting their ideas into practice in the real world.

Generating captions for images has many practical benefits, such as assisting visually impaired individuals and saving costs by automatically labelling the millions of images uploaded online every day. [2] this field combines state-of-the-art models in two main areas of artificial intelligence: natural language processing and computer vision. However, one of the biggest challenges in image captioning is to overfitting of the training data. This is because large datasets, such as the Microsoft Common Objects in dataset, only contain 160k labelled samples. Each top-down architecture constructs (a) a robust image representation, (b) a robust hidden sample to provide a strong foundation for learning expressions, and (c) a language model for a State-LSTM-based representation that captures image semantics. The design is syntactically sound and specifically tailored for generating image captions.

II. LITERATURE SURVEY

It's important to determine time factors, cost-effectiveness, and company strengths before building a tool. Once these requirements are met, pull ahead to Determine the work required for the development of the system. Determine which

operating system the authoring tool is compatible with. Also, consider the platform's flexibility and compatibility with your development environment needs. The developers need extensive external support when they commence working on the tool. Assistance can be obtained from books, websites, and experienced programmers. These elements are considered during the creation of the proposed system. Most of the project development industry considers and thoroughly researches all the requirements needed to implement a project. In literature review. that was important or identify, and also evaluate factors like resource requirements, workers, economics, and business strengths before creating tools and associated designs.

Image classification is a crucial step in the object identification and picture analysis process that leads to the construction of an image sentence. A statement might be the end result of the picture classification process. Many methods for captioning images have been presented Numerous investigations have been conducted to ascertain the most effective method for captioning images. Selecting the best strategy out of all of them is challenging because accuracy and outcomes vary depending on a number of factors. [5] Both new picture captioning techniques developed over the past several decades and conventional ones have undergone ongoing modification to get the best accurate results. Every caption-generating method has advantages and disadvantages of its own. Today's study focuses on merging several methodologies' desirable attributes to increase efficiency .

Numerous complex tasks, including object detection, semantic segmentation, and picture classification, have demonstrated remarkable outcomes when employing multi-layer convolutional neural networks. It is common practice to employ a two-stage approach, particularly for semantic segmentation. In this manner, convolutional networks are trained to provide high-quality local pixel-wise input for the subsequent step, which is frequently a more comprehensive graphical analysis model. To solve the Vanishing Gradient issue, we will employ Long Short-Term Memory (LSTM), a subset of RNNs. The Vanishing Gradients problem is the primary objective of LSTM. The Vanishing Gradients problem is the primary objective of LSTM. Because of its ability to retain data values for extended periods, LSTM is unusual in that it can solve the vanishing gradient problem. An excellent example of this is the image captioning. The task of picture captioning is to provide a sentence-based description of an image given. The image categorization issue and the picture captioning problems are similar in that they both have a larger range of possible answers and require more detail.

III. METHODOLOGY

Three models work together to optimize the process of extracting a caption from a picture, making up the entire system. The types of models are

A. Pattern Recognition Model

This model plays an important role in image captioning. the task of this model is to extract characteristics from a picture to train it. The characteristics of the photos are input when the training starts in this particular model.

As shown in Fig.1, the model employs a VGG16 architecture to effectively extract features from the pictures by combining a sample number which is 3 by filter layers with downsampling layers.Vectors of length 1x4093, which might be used to symbolize the characteristics of the photos, will be the output of the required network.

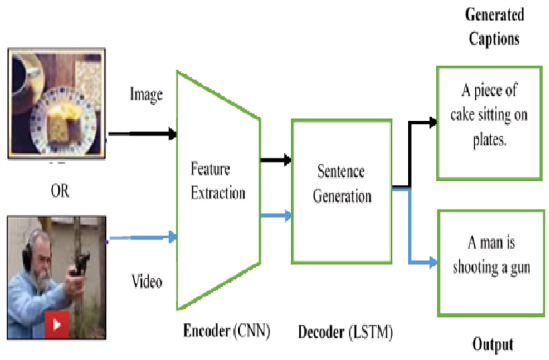

Fig. 1. Methodology of System. [3]

To reduce overfitting, add an output layer of 0.5 to the model. The best value for this metric is 0.5 to 0.8, which represents the probability of this process being discarded.

Taking after the layer, a thick layer is included that basically applies a one-sided activation work to the input part. Rectified Linear Units, or "ReLU," is the activation function that is employed, and 256 is the output space size that is provided. These 256-size vectors are the feature extraction model's outputs, which the decoder model will subsequently employ.

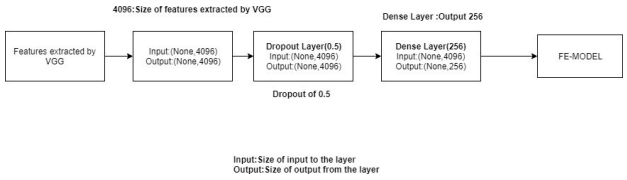

Fig. 2. Pattern Recognition Model [4]

B. Encoder Model

Amid the method of tokenization, the words within the caption are, to begin with changed over to integrability so that the neural organize can prepare them quicker. This is done for each image. Tokenized captions are padded to match the maximum possible sentence length to accommodate all sentences of the same size.

The LSTM Model Layer, or LSTM layer, in the encoder model is learning how words are produced. with the highest likelihood of occurring after a certain word is encountered or

in creating legitimate sentences. ReLU is simply an activation function that is linear in nature, employed and the space is established.

So the generated layer would be easily replaced.and this is

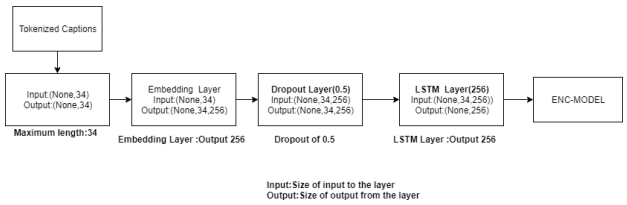

Fig. 3. Encoder Model. [4]

replaced by the Gated Recurrent layer, also to compare the whole model completely, the result of both of these evaluated. The G.R.U layer utilizes the same output-generated space of 265, resulting in the encoder component being the sole differentiating factor between the two models. The output of the LSTM layer serves as the output for the encoder layer.

C. Decoder Model

It combines the pattern recognition model and model and the Converter model to provide desired results, in the decoder model, as depicted in Figure 3. The input is provided by both the model extractor and the encoder model, which produce the vectors whose approximate size is 256 only. The values of the samples are connected using the activation function ”ReLU” as an intensity method. Additional thickness layers added to the decision model use mass points as the output source. For each desired number, softmax activation essentially produces a word. The desired single term emerges as the output, of decoding stratum. The enter elements encompass the image and the input collection.

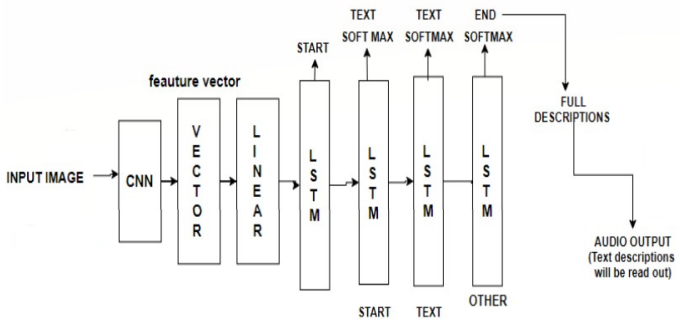

Fig. 4. Final Decoder Model. [6]

IV. TECHNOLOGY USED

A. Python

Python is an interpreted general-purpose programming language. Guido van Rossum created Python, Its design philosophy is notable for its extensive use of whitespace and its high emphasis on code readability. It’s based on object- approach and linguistic constructions are intended to help developers create clear, code for small and large applications alike. Python has variable typing and garbage collection as well. Functional, legal, and object-oriented design are among the programming paradigms it supports.

B. React

React is a development project that provides a component-based design that allows developers to create reusable components that can be quickly updated and executed when data changes. React’s virtual Document Object Model (DOM) reduces the need to modify the browser DOM by changing only the elements that are actually needed. Developers determine what the user interface should look like based on the state of the application at any given moment. This approach increases the readability and maintainability of the code.

C. Jupyter Notebook

Jupyter Notebook stands as an accessible web-based platform enabling users to generate and distribute documents featuring live code, mathematical equations, graphical visualizations, and explanatory annotations. Its versatile utility spans various domains such as data manipulation, numerical analysis, statistical inference, data representation, machine learning, and more.

V. DATASET SPECIFICATIONS

The intention of this undertaking is for the captions of the input photographs to be predicted and generated. [**?**] Datasets are used for this purpose. The input pictures are evaluated to generate correct captions of the photographs. The dataset incorporates 8k pictures with 5 labels per photograph. For installation, capabilities are extracted from pictures and textual labels. The functions are then combined to predict the subsequent word within the tag. CNN is taken for recognition of the images, and for text generation LSTMs are used instead of CNN. we use BLEU for evaluating the efficiency and performance of the overall model that is used.

A. Flickr8K Dataset

Flickr8k is a widely used program for linking images with captions. The collection includes 8k pictures, each with five unique titles, capturing a variety of scenes and landscapes. Importantly, it deliberately does not include images featuring prominent people or visible signs to ensure it is applicable across multiple contexts. The training set contains 6000 images, and the development set and testing set contains 1000 images each to facilitate evaluation.

VI. ANALYSIS

The above picture is an examples used for testing.Figure 4. I tried both VGG+LSTM and VGG+GRU approaches on different photos. The model was trained on Google Colab with a single Tesla K80 GPU with 12 GB GDDR5 VRAM. For LSTM and GRU, this took approximately 13 and 10 minutes per epoch, respectively. This is because GRU performs fewer operations than LSTM. Although the LSTM model had lower

loss than GRU, users are free to choose the model that best suits their needs, whether it is the highest accuracy or the fastest processing time. Compared to LSTM, GRU often learns faster with less training data and is easier to use.

Due to its complexity, demanding additional time for training and sentence formulation. the LSTM model performs marginally better than the GRU in most cases. By training on a greater number of photos, a larger dataset is also anticipated to improve performance. With our integrated text-to-speech technology, visually impaired persons can also benefit substantially and have a greater feel of their surroundings due to the generated image captions' notable accuracy.

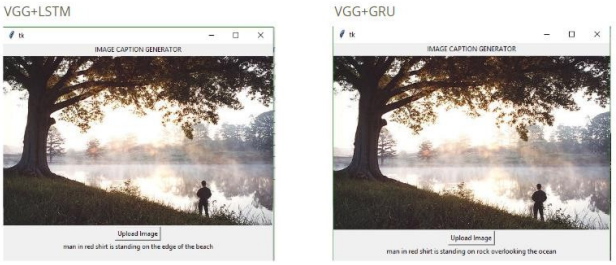

Fig. 5. Final Decoder Model. [4]

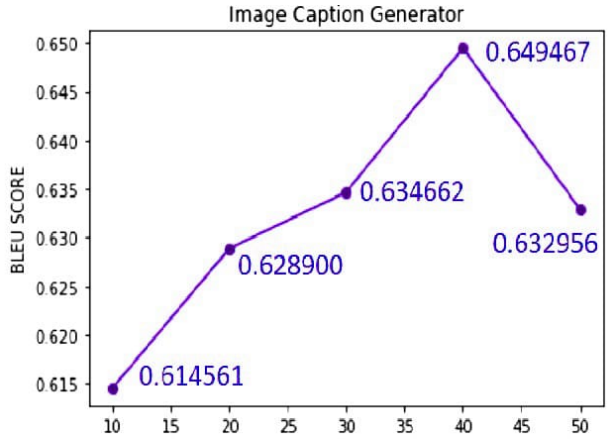

Fig. 6. BLEU Score comparison [8]

VII. CONCLUSION

Employing a CNN highlight extraction demonstrates to encode an picture into a vector representation and an RNN decoder demonstrates to create comparing sentences based on the picture highlights learned, displayed a machine learning model demonstration that has a propensity to spontaneously produce image descriptions. to assist outwardly disabled individuals superior get it their situations. We have too compare diverse Utilizing models to analyze the effects of each blocks on the captioning era and possess a significant impact. too illustrated different utilize cases on our framework. The discoveries illustrate that, in spite of the fact that We utilize the Flickr 8K dataset to prepare our demonstration. is very little and has less diverse picture sorts. The Flickr8K datasets will be utilized to prepare our calculation, which is able to empower us to deliver more precise expectations. Altering the hyperparameters, such as clump estimate, number of ages, learning rate, etc., and deciding how each one influences our demonstrate an illustration of the model.

VIII. FUTURE WORK

Since our model is not flawless, it occasionally produces inaccurate captions. We will be creating models in the following stage that employ Inceptionv3 as the feature extractor rather than VGG. The four models that were so produced. This will assist us in our analysis of CNN's impact across the network as a whole.

At the moment, to select those words with the probability chance is high to generate next word in sequence in greedy approach. Alternatively, beam search chooses a set of words with the highest probability and runs parallel searches across every sequence. We might be able to work on the result accuracy of the model and improve it with the help of the given model.

REFERENCES

[1] Megha J Panicker, Vikas Upadhayay, Gunjan Sethiand Vrinda Mathur, "Image Caption Generator", International Journal of Innovative Technology and Exploring Engineering (IJITEE), vol. 10, no. 3, pp. 87–92, Jan. 2021, doi: 10.35940/ijitee.C8383.0110321.

[2] Sharma, Grishma and Kalena, Priyanka and Malde, Nishi and Nair, Aromal and Parkar, Saurabh, Visual Image Caption Generator Using Deep Learning (April 8, 2019). 2nd International Conference on Advances in Science Technology (ICAST) 2019 on 8th, 9th April 2019 by K J Somaiya Institute of Engineering Information Technology, Mumbai, India, Available at SSRN: https://ssrn.com/abstract=3368837 or http://dx.doi.org/10.2139/ssrn.3368837

[3] Image Caption Generator using deep learning with Flickr Dataset ", International Journal of Science Engineering Development Research (www.ijrti.org), ISSN:2455-2631, Vol.7, Issue 8, page no.1145 - 1152, August-2022, Available :http://www.ijrti.org/papers/IJRTI2208183.pdf

[4] Sharma, Grishma Kalena, Priyanka Malde, Nishi Nair, Aromal Parkar, Saurabh. (2019). Visual Image Caption Generator Using Deep Learning. SSRN Electronic Journal. 10.2139/ssrn.3368837. Magnetics Japan, p. 301, 1982].

[5] Y. Feng and M. Lapata, "Automatic Caption Generation for News Images," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 35, no. 4, pp. 797-812, April 2013, doi: 10.1109/TPAMI.2012.118. keywords: Visualization;Humans;Databases;Vocabulary;Probabilistic logic;Data models;Noise measurement;Caption generation;image annotation;summarization;topic models,

[6] Vinyals, O., Toshev, A., Bengio, S., Erhan, D. (2014, November 17). Show and Tell: a neural Image caption generator. arXiv.org. https://arxiv.org/abs/1411.4555 Vinyal, Alexander Toshev, Samy Bengio, Dumitru Erhan, IEEE (2015).

[7] Palak Kabra, Mihir Gharat, Dhiraj Jha, Shailesh Sangle."Image Caption Generator Using Deep Learning", Volume 10, Issue X, International Journal for Research in Applied Science and Engineering Technology (IJRASET) Page No: 621-626, ISSN : 2321-9653, www.ijraset.com

[8] Liu, Chang Wang, Changhu Sun, Fuchun Rui, Yong. (2016). Image2Text: A Multimodal Image Captioner. 746-748. 10.1145/2964284.2973831.

[9] P. Mathur, A. Gill, A. Yadav, A. Mishra and N. K. Bansode, "Camera2Caption: A real-time image caption generator," 2017 International Conference on Computational Intelligence in Data Science(ICCIDS), Chennai, India, 2017, pp. 1-6, doi: 10.1109/ICCIDS.2017.8272660. keywords: Computational modeling;Decoding;Training;Mathematical model;Real-time systems;Feature extraction;Visualization,

[10] Rennie, Steven J., Etienne Marcheret, Youssef Mroueh, Jerret Ross and Vaibhava Goel. “Self-Critical Sequence Training for Image Captioning.” 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016): 1179-1195.

[11] Fang, H., Gupta, S., Iandola, F., Srivastava, R., Deng, L., Dollár, P., Gao, J., He, X., Mitchell, M., Platt, J. C., Zitnick, C. L., Zweig, G. (2014, November 18). From captions to visual concepts and back. arXiv.org. https://arxiv.org/abs/1411.4952

[12] J. Donahue et al., ”Long-term recurrent convolutional networks for visual recognition and description,” 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 2015, pp. 2625-2634, doi: 10.1109/CVPR.2015.7298878. keywords: Visualization;Computer architecture;Computational modeling;Data models;Logic gates;Image recognition;Microprocessors,

[13] C. Szegedy et al., ”Going deeper with convolutions,” 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 2015, pp. 1-9, doi: 10.1109/CVPR.2015.7298594. keywords: Computer architecture;Convolutional codes;Sparse matrices;Neural networks;Visualization;Object detection;Computer vision

[14] Simonyan, K., Zisserman, A. (2014, September 4). Very deep convolutional networks for Large-Scale image recognition. arXiv.org. https://arxiv.org/abs/1409.1556v6

[15] SUN, Le XU, Bin LU, Zhenyu. (2022). Hyperspectral Image Classification Based on A Multi-Scale Weighted Kernel Network. Chinese Journal of Electronics. 31. 832-843. 10.1049/cje.2021.00.130.

[16] Hirose, Akira Yoshida, Shotaro. (2012). Generalization Characteristics of Complex-Valued Feedforward Neural Networks in Relation to Signal Coherence. IEEE Transactions on Neural Networks. 23. 541-551. 10.1109/TNNLS.2012.2183613.

[17] Walke, J. (2013). ReactJS by Facebook: A-declarative, efficient, and flexible JavaScript library for building user interfaces. IEEE Software.