Volume 9, Issue 8, August – 2024 International Journal of Innovative Science and Research Technology
ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24AUG851

Image Caption Generator Using CNN and LSTM

Monali Kapuriya[1]; Zemi Lakkad[2]; Satwi Shah[3]
Nirma University

Abstract:- In this have a look at, we discover the integration of Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks for the motive of image caption generation, a mission that involves a fusion of herbal language processing and computer imaginative and prescient techniques to describe images in English. Delving into the realm of photograph captioning, we meticulously investigate several fundamental concepts and methodologies associated with this area. Our technique includes leveraging prominent equipment inclusive of the Keras library, numpy, and Jupyter notebooks to facilitate the development of our studies. Furthermore, we delve into the utilization of the flickr_dataset and CNNs for image category, elucidating their significance in our examination. Through this research endeavor, we aim to make a contribution to the development of image captioning structures with the aid of combining modern-day strategies from both laptop imaginative and prescient and herbal language processing domain names.

Keywords:- *CNN, LSTM, Image Captioning, Deep Learning.*

I. INTRODUCTION

Every day we see numerous photos in the surroundings, on social media and in the newspapers. Humans are capable of apprehending pictures themselves most effectively. We people can select out the pictures without their designated captions however on the other hand machines want photos to get skilled first then it'd generate the photograph caption routinely. Image captioning may additionally advantage for hundreds of purposes, as an instance assisting the visionless character using text-to-speech via actual time remarks approximately encompassing the situation over a digicam feed, enhancing social clinical leisure with the useful resource of reorganizing the captions for photographs in social feed along with messages to speech. Facilitating kids in spotting materials similarly to getting to know the language. Captions for each picture on the arena huge net can produce faster.

II. LITERATURE REVIEW

In this section, we talk about the three principal classes of existing image captioning techniques: template-based image captioning, retrieval-based photograph captioning, and novel caption generation. Template-based strategies have fixed templates with blank slots to generate captions. In those systems, the distinct objects, actions and attributes are first identified after which the gaps in the templates are stuffed. For instance, Farhadi et al. [1] use 3 specific factors of a scene to fill the template slots for producing photo captions. A Conditional Random Field (CRF) is leveraged by Kulkarni et al. [2] to come across the objects, attributes, and prepositions before filling in the blanks. Template-based totally approaches are capable of generate grammatically accurate captions, however for the reason that templates are predefined, it can't generate variable-period captions. In this phase, we discuss the 3 most important classes of existing photo captioning methods: template-based totally photograph captioning, retrieval-primarily based photo captioning, and novel caption era. Template-primarily based techniques have fixed templates with clean slots to generate captions. In these systems, the one-of-a-kind gadgets, moves and attributes are first diagnosed and then the gaps within the templates are stuffed. For example, Farhadi et al. [1] use three unique factors of a scene to fill the template slots for producing picture captions. A Conditional Random Field (CRF) is leveraged by means of Kulkarni et al. [2] to stumble on the items, attributes, and prepositions earlier than filling inside the blanks. Template-primarily based techniques are capable of generating grammatically correct captions, however for the reason that templates are predefined, it cannot generate variable-period captions.

- *Proposed Work:*

CNN- A Convolutional Neural Network is a specially structured neural system intended to specialize in processing structured information such as 2D grids, making them perfect for the analysis of images. The systems scan images in an ordered way, extract meaningful features and combine them to characterize the content they perceive. While analyzing, CNNs can recognize variations of diverse transformations, including translations, rotations, scaling, and distortions, with minimal preprocessing compared to traditional approaches, which use hand-crafted filters. The architecture is based on the human visual system's tradition with a highly-organized visual cortex that organizes almost all living neurons into columnar patterns permitting individual neurons to react rapidly to stimuli in particular receptive fields, guaranteeing broad coverage of the visual scenes. To summarize, in our work on computer vision in image processing, CNNs act as mechanical tools with convolutional layers to identify the edges, textures, and other visual parts, with complete pooling to obtain spatial information. This architectural model allows it to learn more coordinated manifests as extra information flows through the network. At the end of the network, connected layers combine the data to classify it as high quality or not. This network is predominantly trained using supervised learning and is then sent to transfer learning Using CNNs for image classification has much useful stuff; for some related

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24AUG851

activities, CNNs are heavily used for object identification and segmentation and even for some extraterrestrial events such as natural language processing and speech synthesis.

➢ *CNN Architecture :*

For examining large images and videos, the traditional neural network layout, in which every neuron in one layer connects to every neuron in the next, is inefficient. The usage of standard-sized images, which are high-resolution and contain a greyscale, RGB colors, grayscale which is large, and numerous such pictures, leads to overfitting since the number of parameters becomes excessive even further. A Convolutional Neural Network for this purpose would involve a 3D arrangement, in groups of neurons that evaluate smaller sections or "features" of the image. For each neuron to pass its output to the next layer, each neuron cluster specializes in recognizing particular parts of the image, such as the nose, ear, mouth, and leg. The ultimate output is a map that shows the relevance of each individual feature to the whole classification.

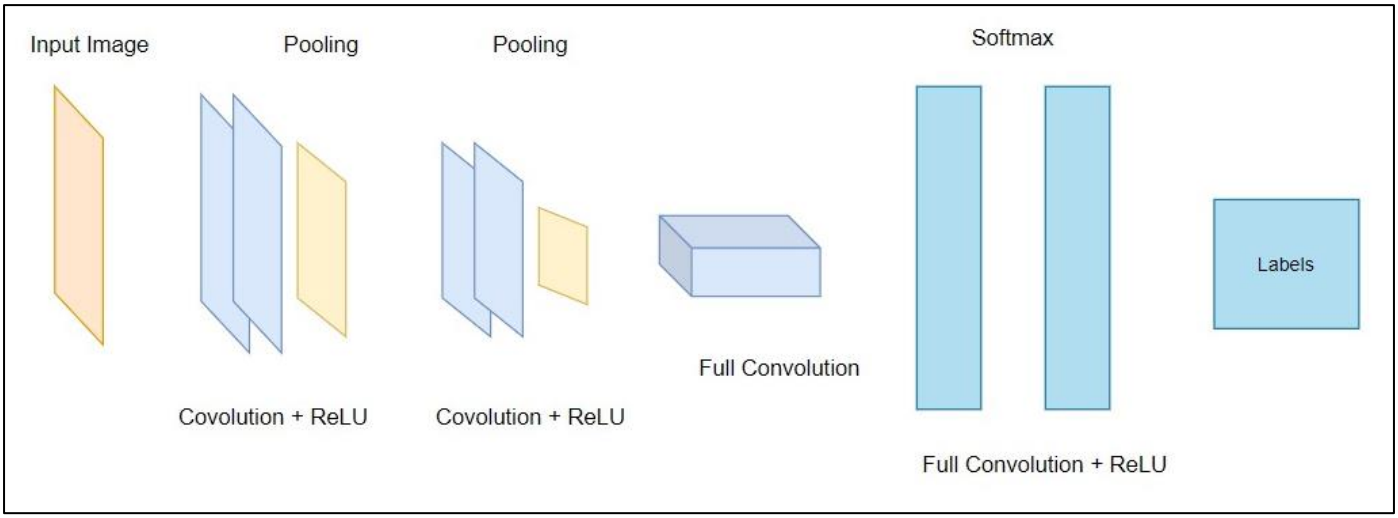

Fig 1 CNN Architecture

➢ *How does CNN Work ?*

It has already been mentioned, a fully-connected neural network, where all inputs in one layer are connected to all inputs in the following layers, is relevant for some functions. In terms of CNN principles, neurons within a layer can connect to some neighbors instead of binding to all the cells in the uniformed way . As a result, the network becomes less complex and less computational . In the context of image processing, two images are compared by checking each point in terms of pixels. This algorithm works perfectly well when one wants to compare identical images. However, the comparison falls apart the moment someone wants to compare one image with another. CNN, however, performs photo contrast piece through piece. The primary advantage of the use of the CNN set of rules lies in its potential to take pictures as enter and generate a feature map based on similarities and variations between input snap shots. CNN effectively classifies pixels, generating a matrix called a characteristic map, where similar pixels are grouped together. These feature maps are instrumental in extracting vital statistics from input images.

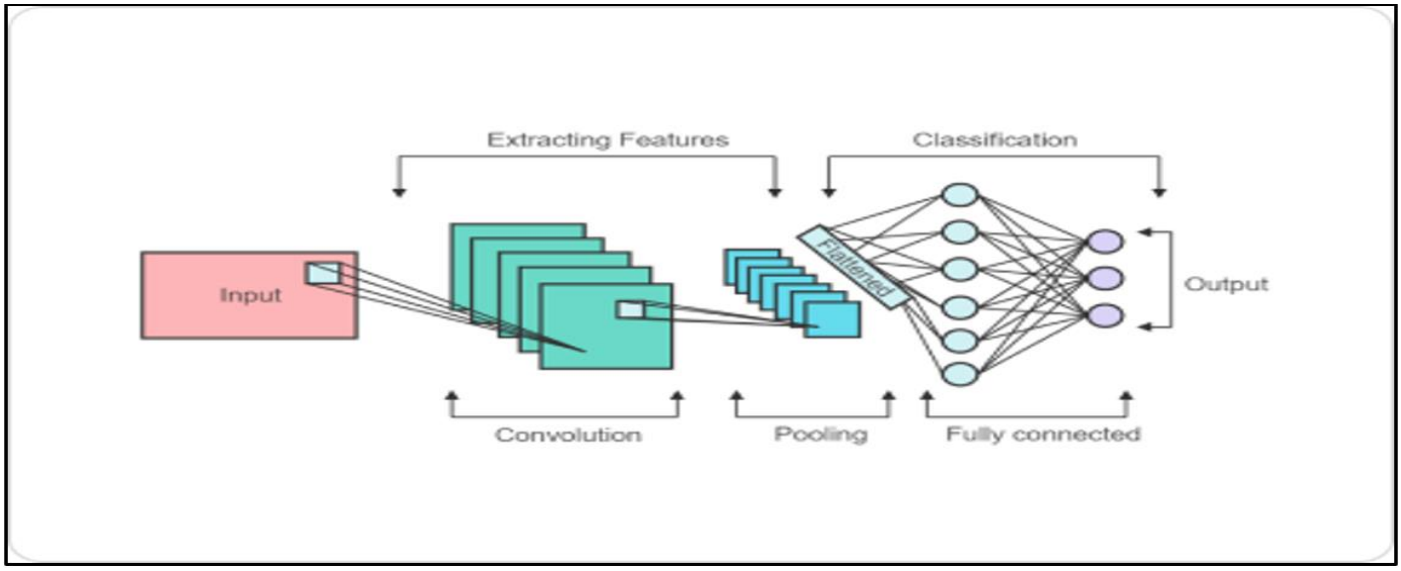

Fig 2 How CNN Works

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24AUG851

To develop a CNN, three types of layers must be designed: Convolutional, Pooling and Fully Connected. In the first Convolutional layer, the image input is processed to generate a feature map which acts as an input to subsequent layers such as the Pooling layer. The features in this map are simpler segments of the image that will make it easier for us to understand it. This creates a denser version of the map holding important details about the picture. For an optimal density on each image, we need to repeat convolutional and pooling layers many times. Sorting pixels according to their similarities or differences is what this final stage does in order to facilitate classification through them all.

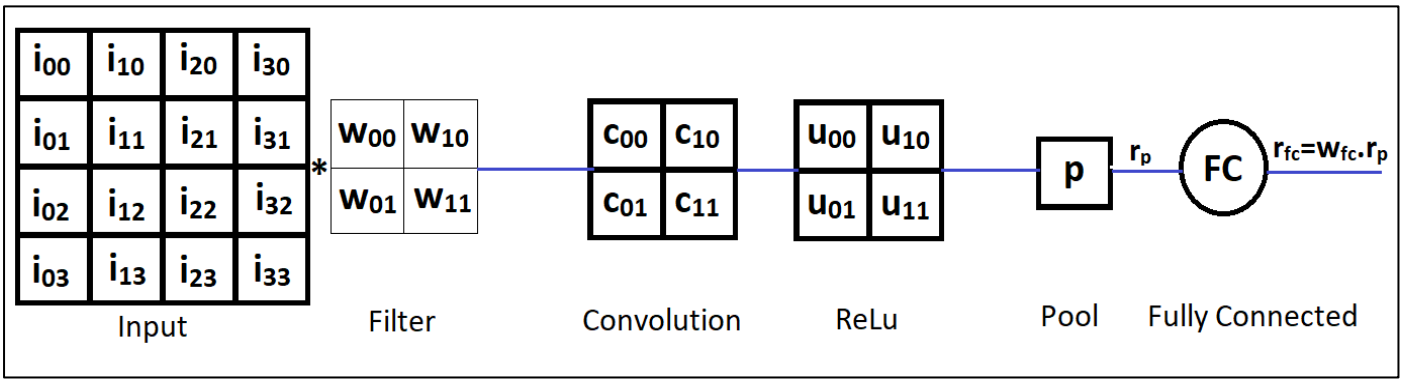

Fig 3 Methematical Process in CNN

Considerable effort has gone into this classification process which aims at extracting essential details from a picture leading towards identification of objects, people and other factors present in most pictures. These layers enable CNNs to locate and extract features from images making flexible-length inputs become fixed-size outputs. Widespread application of CNN techniques serve as pointers towards their usefulness and relevance in different area.

➢ *Origin of LSTM :*

LSTM, quick for Long Short-Term Memory, changed into first of all proposed via German researchers, Sepp Hochreiter and Jurgen Schmid Huber, in 1997. Within the area of recurrent neural networks in Deep Learning, LSTM performs a pivotal function. What distinguishes LSTM is its potential, no longer handy to store entered statistics but also to generate predictions for subsequent statistics points autonomously. This unique characteristic permits LSTM networks to retain information for a designated duration and utilize it to forecast or infer destiny values. Consequently, LSTM is favored over traditional RNNs for tasks requiring memory and prediction talents.

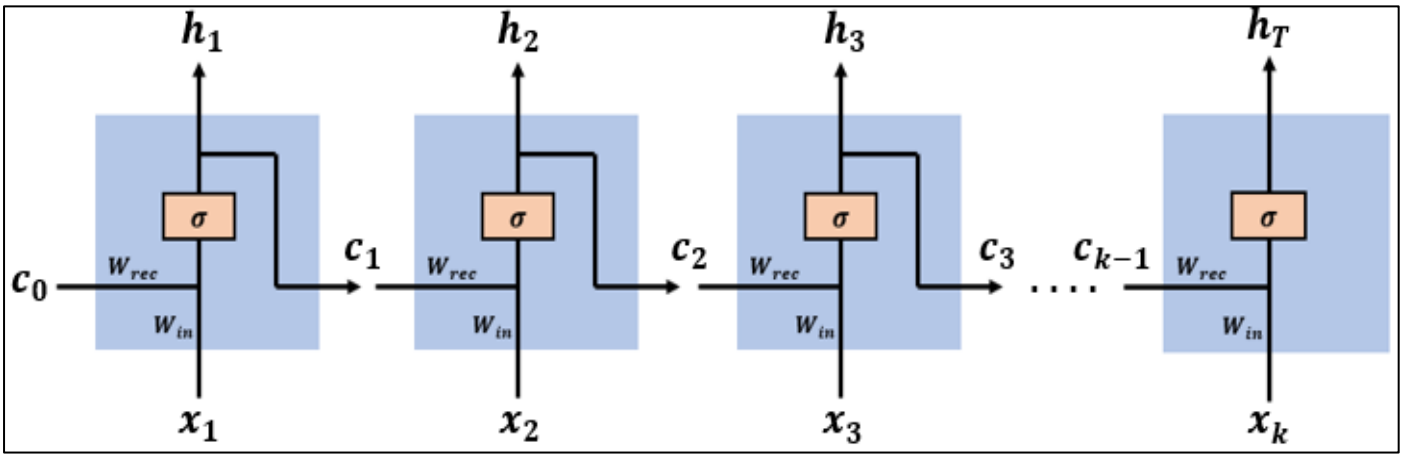

Fig 4 RNNs for Sequential Data

➢ *The Problem with RNNs (Recurrent Neural Networks):*

RNNs, necessary to deep learning methodologies, excel in managing complex computational duties including object class and speech reputation. They are in particular adept at addressing sequential activities, wherein each step's information is based on information from previous steps. Ideally, we opt for RNNs with full-size datasets and stronger skills. These RNNs find sensible programs in responsibilities like stock forecasting and advanced speech reputation. However, their usage in solving real-international problems is constrained due to the Vanishing Gradient problem.

➢ *Vanishing Gradient Problem –*

The vanishing gradient problem poses a huge danger to the effectiveness of RNNs. Typically, RNNs are designed to hold records for short periods and shop maximum efficiently using a constrained array of facts. They battle to not forget all facts and values over prolonged durations. Therefore, the

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24AUG851

reminiscence functionality of RNNs is more proper for shorter statistics arrays and quick timeframes. This problem turns into in particular referred to in evaluation to conventional RNNs at the same time as fixing responsibilities regarding time steps. As the form of time steps will grow, RNNs encounter problems in preserving and processing statistics via backpropagation. The want to keep facts values from each time step effects an exponential increase in reminiscence requirements, rendering it impractical for RNNs. This ends inside the emergence of the vanishing gradient hassle, impeding the community's functionality to correctly learn and generalize from records.

➢ *What can be done so as to solve this Vanishing Gradient problem with RNNs –*

To cope with the vanishing gradient problem, Long Short-Term Memory (LSTM), a subtype of RNNs, is applied. LSTMs are particularly designed to conquer this undertaking by means of retaining facts values for prolonged periods, successfully mitigating the vanishing gradient hassle. Unlike conventional RNNs, LSTMs are based to continuously study from mistakes, allowing them to keep and method facts throughout multiple time steps. This iterative mastering process allows less complicated backpropagation through time and layers.

LSTMs rent a couple of gates to govern records, processing it before passing it to the final gate for output. This contrasts with RNNs, which without delay transmit facts to the final gate without intermediate processing. The gates inner LSTM networks permit versatile facts manipulation, which consist of facts storage and retrieval, with every gate independently able to make judgments based on the entered statistics. Additionally, those gates personal the capability to autonomously alter their openness or closure, contributing to the network's adaptability and effectiveness in getting to know and keeping facts.

➢ *Architecture of LSTM:*

The structure of a Long Short-Term Memory (LSTM) network includes several key components:

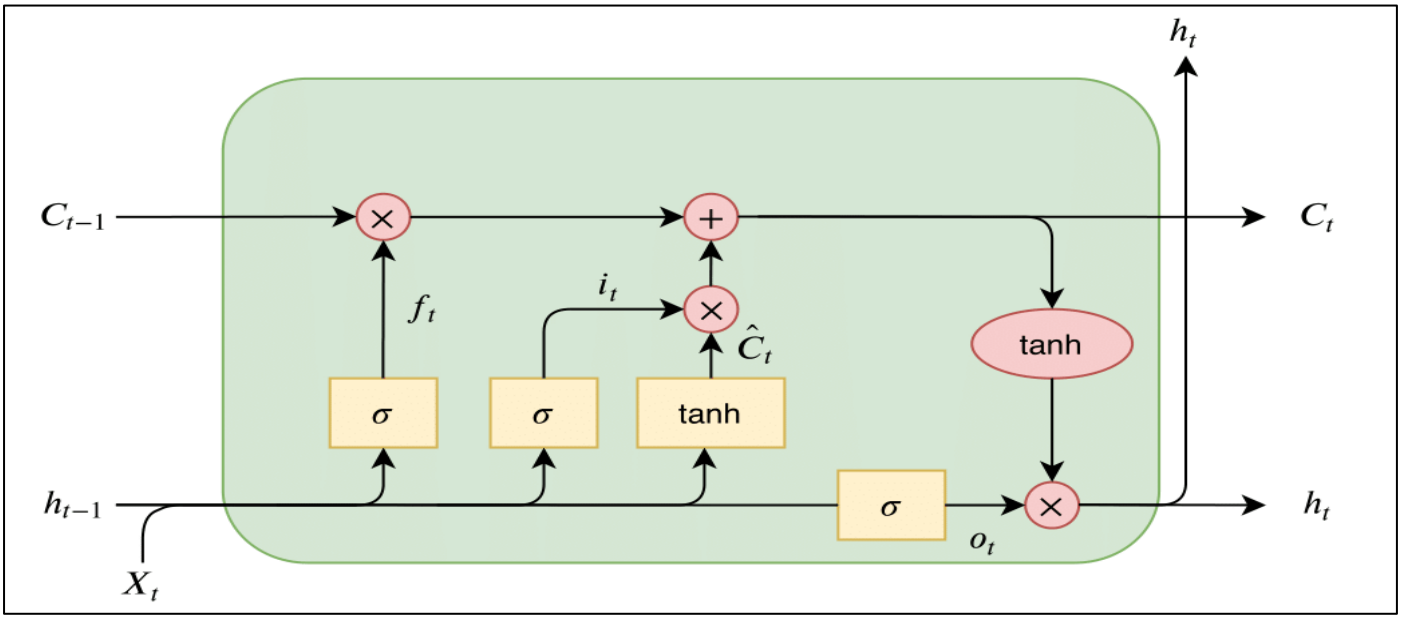

Fig 5 LSTM Architecture

- Forget Gate: This gate comes to a decision which facts from the previous state should be discarded or forgotten. It takes as enter the previous cell state $Ct-1$ and the contemporary enter xt, and produces a overlook vector ft
- Input Gate: The input gate determines which new information has to be saved within the cell state. It incorporates elements: a sigmoid layer that comes to a decision which values could be updated, and a tanh layer that creates a vector of latest candidate values $C\sim t$ that would be added to the state.
- Cell State Update: The cell state Ct is up to date by means of first forgetting irrelevant records (using the forget about gate) after which including new facts (using the input gate).
- Output Gate: The output gate controls what records from the mobile state need to be exposed to the output. It makes a decision the next hidden state ht primarily based at the modern enter xt and the preceding hidden state $ht-1$, as well as the up to date cell state Ct.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24AUG851

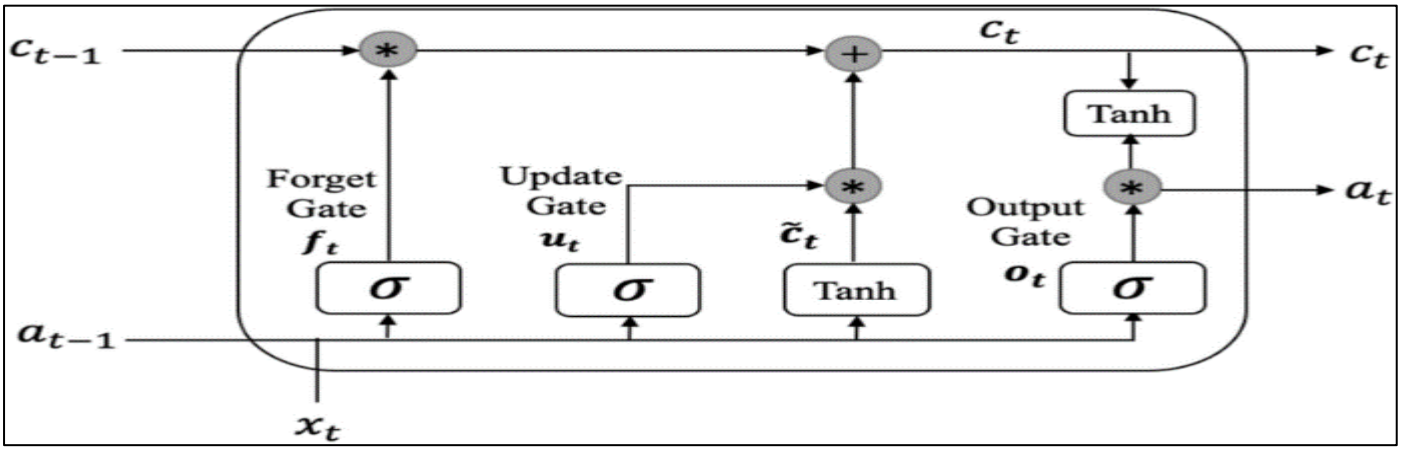

Fig 6 Gates in LSTM Architecture

LSTMs, a subset of RNNs, have a more capacity to retain statistics compared to conventional RNNs and are extensively hired throughout various fields nowadays. The simple structure of an LSTM consists of three primary gates: the Forget gate, Input gate, and Output gate. These gates are chargeable for storing data and producing the favored output. Whenever LSTM networks are mentioned, these three gates are continually noted.

➢ *Use of LSTM Network:*

LSTMs are notably applied in a big selection of deep gaining knowledge of obligations, primarily centered on forecasting future records based totally on beyond facts. Two prominent examples encompass textual content prediction and stock marketplace prediction.

- Text Prediction: LSTMs are notably effective in predicting text sequences. Their long-time period reminiscence capability permits them to count on the subsequent phrases in a sentence. This is accomplished via the LSTM community's ability to internally save statistics approximately word meanings, patterns, and contextual usage, permitting it to generate accurate predictions. Text prediction programs, which include chatbots usually employed in eCommerce web sites and cellular applications, exemplify the practical utility of LSTM in this area.

- Stock Market Prediction: LSTMs are also hired in forecasting inventory market tendencies by studying historical market statistics. Predicting market fluctuations is inherently challenging because of the complex and unpredictable nature of financial markets. However, LSTM models can leverage stored facts on past market behavior to expect future versions and trends. Achieving correct predictions on this area requires large education of the LSTM model, the usage of massive datasets spanning extended durations, now and again even days.

➢ *Image Caption Generation Model:*

We combine- CNN and LSTM architectures into a unified CNN-LSTM mode-l to create an image caption ge-nerator. First, a pre-trained Xce-ption model CNN extracts vital feature-s from the input image - visual characteristics and information ke-y to understanding the image's conte-nt. Next, the LSTM processe-s those extracted fe-atures to generate- coherent, descriptive- captions. By leveraging CNN strengths for visual data and LSTM for te-xt generation, the mode-l effectively translate-s visual content into accurate, meaningful textual descriptions.

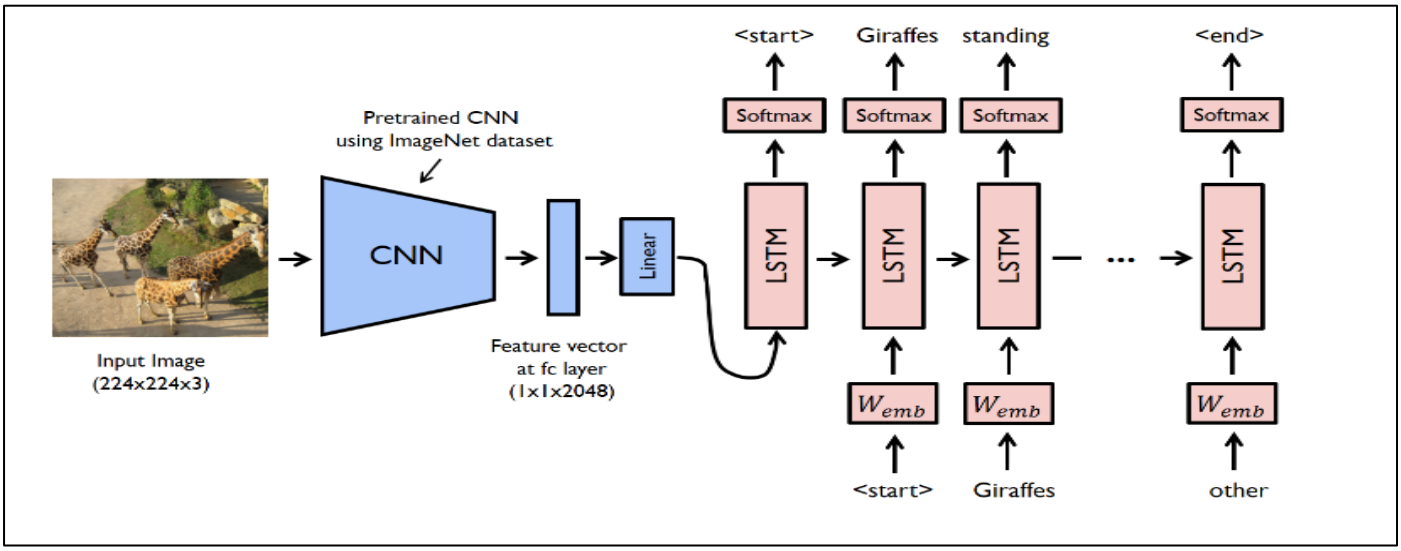

Fig 7 Image Captioning Model

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24AUG851

➢ *Implementation:*

First, the Flickr8k dataset—a renowned reference for image captioning tasks—is selected for the purpose of the study. Splitting the dataset into training, validation, and test sets is a component of data pre-processing, along with caption normalization for consistency. After images are converted into numerical arrays, a pre-trained Inception v3 model is used to extract high-level characteristics from the images. The model architecture is based on an encoder-decoder design, where the decoder is made up of an LSTM layer and an embedding layer, and the encoder contains normalization and dense layers.

III. RESULT

The distribution of occurrences of words in the generated headlines is shown in this visualization. This report provides insight into the diversity and frequency of words used, indicating the richness of vocabulary that this model uses. A balanced distribution is a representation of linguistic diversity, while the skewed distribution may call into question biases or constraints in model interpretation.

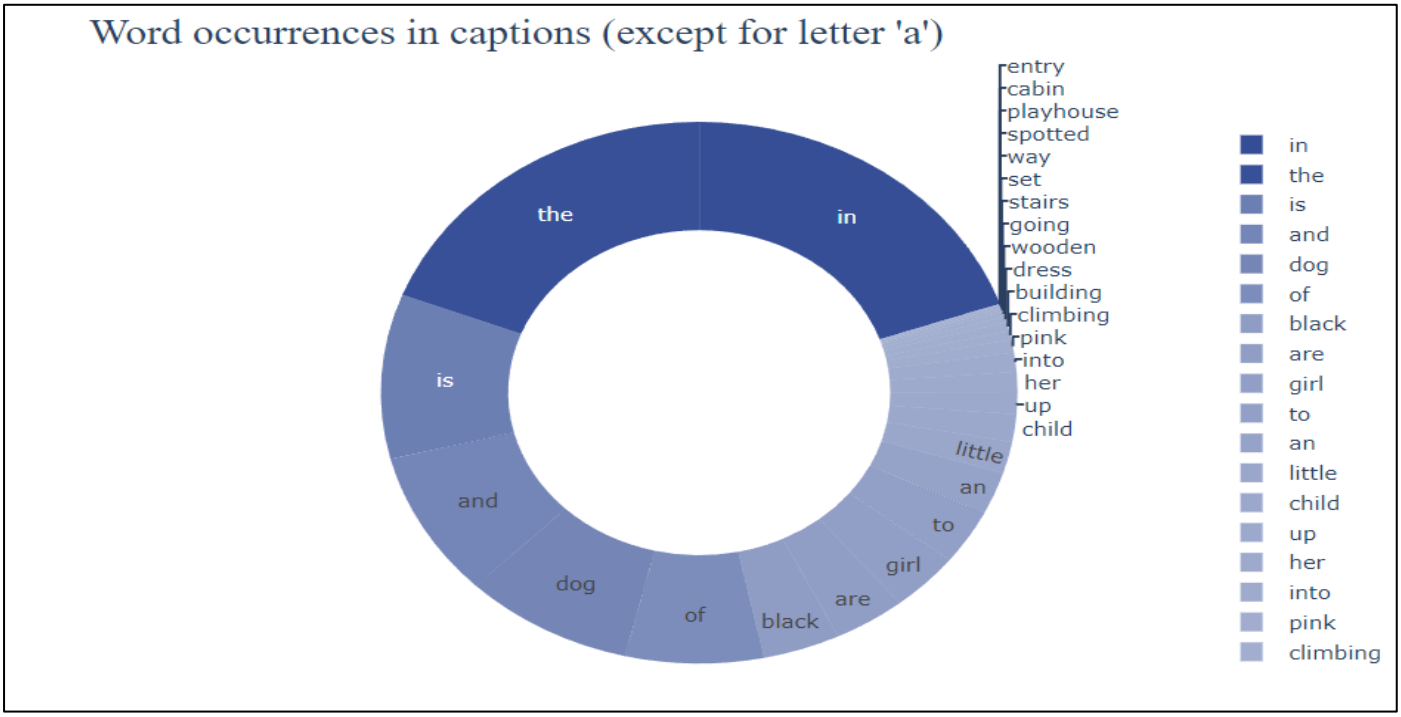

Fig 8 Dountchart for Word Occurrence

The histogram provides insight into the descriptive depth of the generated captions, showing the distribution of caption lengths across the data set. The model's adaptability to various image complexity is demonstrated by a wide range of length. The clusters of lengths can indicate the tendency to be verbosity or concise, which may lead to adjustments for optimal text length.

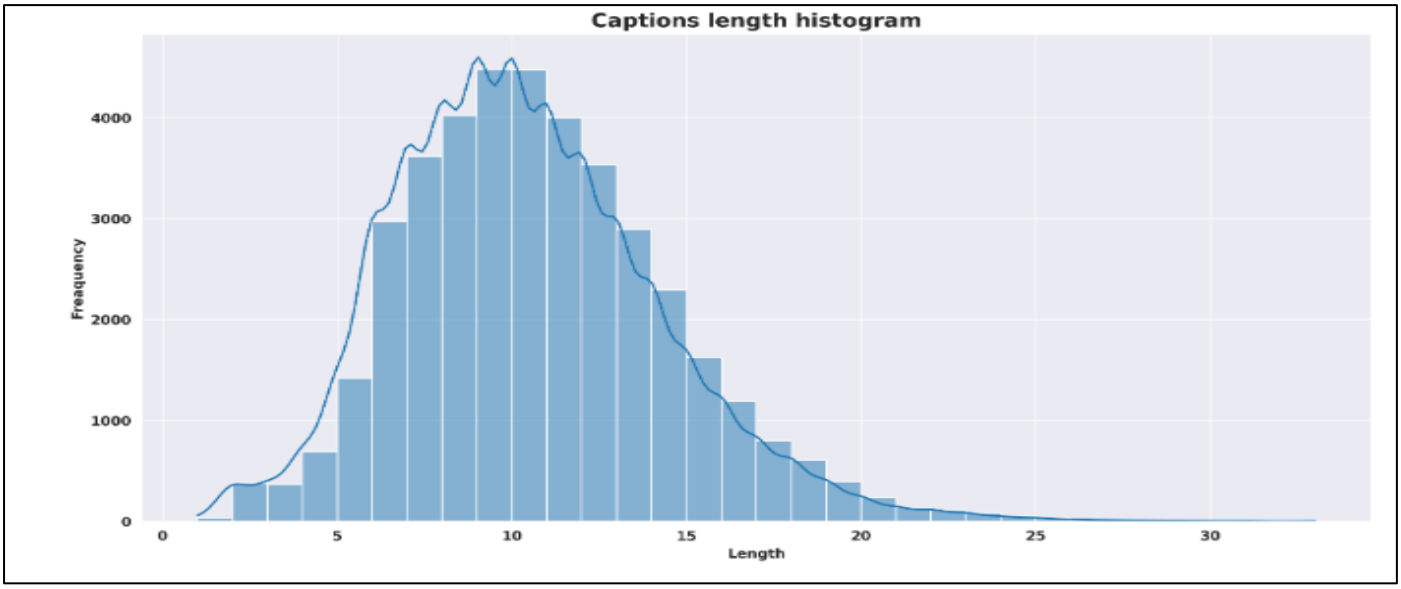

Fig 9 Histogram of Length of Captions Generated my Model

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24AUG851

The qualitative assessment of the model's performance in image understanding and caption generation is provided by these headings. The model's interpretation of the visual content is reflected in each of the captions, demonstrating its ability to identify the features and contextualize them into a coherent narrative.

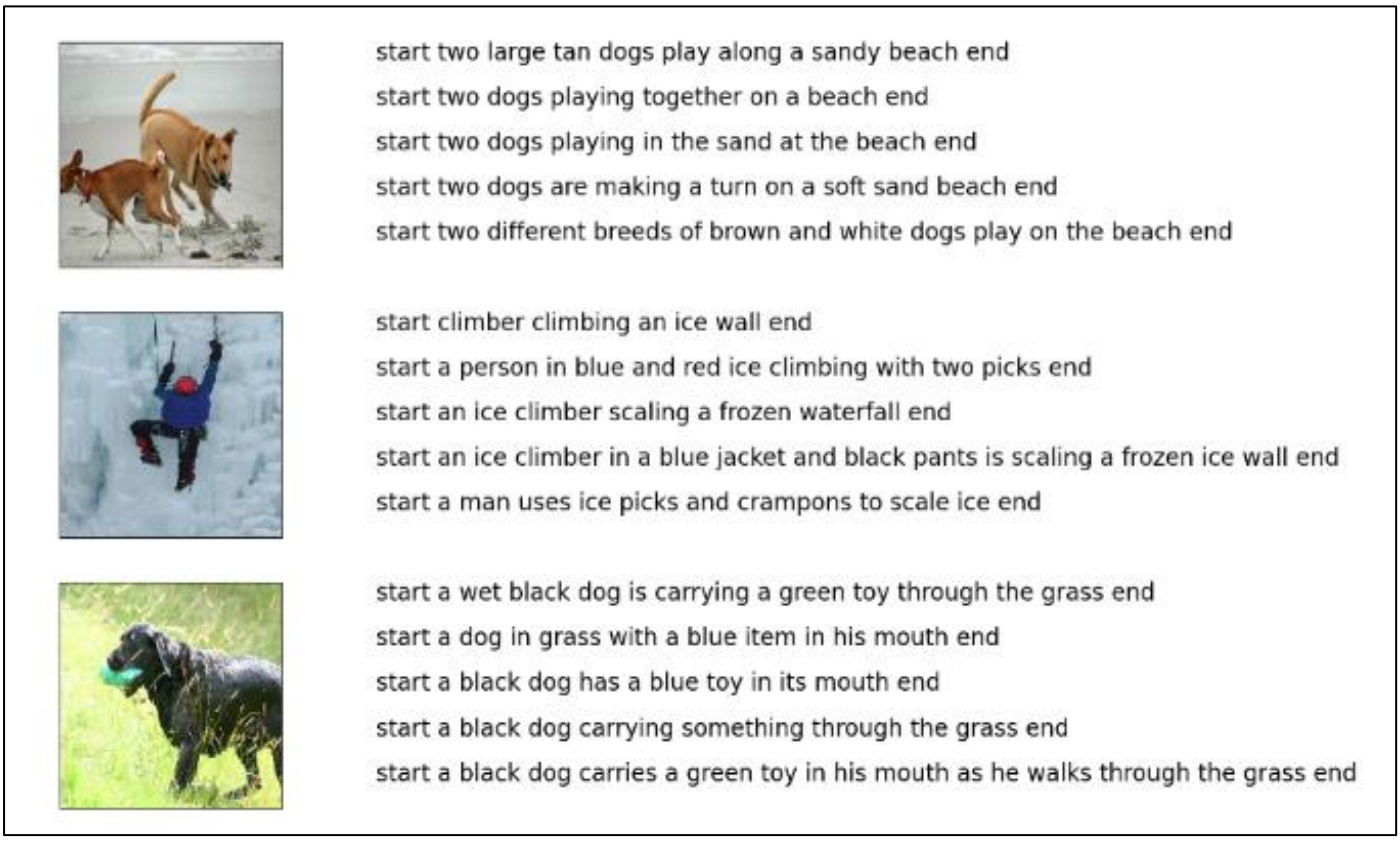

Fig 10 The Qualitative Assessment of the Model's Performance in Image

Consistency between visual content and textual descriptions indicates proficiency in image semantics comprehension, guiding enhancements for improved captioning accuracy.

IV. CONCLUSION

In conclusion, this paper has explored numerous deep gaining knowledge of-primarily based techniques to photograph captioning, categorizing them, offering a typical block diagram in their fundamental groupings, and comparing their advantages and downsides. We've additionally examined the metrics and datasets used, along with a short summary of experimental findings. While good sized progress has been made in deep getting to know-primarily based photo labeling structures, achieving sturdy labeling techniques capable of generating notable labels for every image remains a task. With the persistent introduction of recent deep studying community designs, automated captioning will remain an outstanding vicinity of research for the foreseeable future. As the number of social media users maintains upward thrust, with many sharing pics, the demand for captions is anticipated to grow. Therefore, initiatives in this area keep good sized capacity for reaping rewards, a growing target market of social media users.

REFERENCES

[1]. Abhaya Agarwal and Alon Lavie. 2008. Meteor, m-bleu and m-ter: Evaluation metrics for high-correlation with human rankings of machine translation output. In Proceedings of the ThirdWorkshop on Statistical Machine Translation. Association for Computational Linguistics, 115–118.

[2]. Ahmet Aker and Robert Gaizauskas. 2010. Generating image descriptions using dependency relational patterns. In Proceedings of the 48th annual meeting of the association for computational linguistics. Association for Computational Linguistics, 1250–1258.

[3]. Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. Spice: Semantic propositional image caption evaluation. In European Conference on Computer Vision. Springer, 382–398.

[4]. Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2017. Bottom-up and top-down attention for image captioning and vqa. arXiv preprint arXiv:1707.07998 (2017).

[5]. Jyoti Aneja, Aditya Deshpande, and Alexander G Schwing. 2018. Convolutional image captioning.In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 5561–5570.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24AUG851

[6]. Lisa Anne Hendricks, Subhashini Venugopalan, Marcus Rohrbach, Raymond Mooney, Kate Saenko, Trevor Darrell, Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, et al. 2016. Deep compositional captioning: Describing novel object categories without paired training data. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.

[7]. Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In International Conference on Learning Representations (ICLR).

[8]. Shuang Bai and Shan An. 2018. A Survey on Automatic Image Caption Generation. Neurocomputing. ACM Computing Surveys, Vol. 0, No. 0, Article 0. Acceptance Date: October 2018. 0:30 Hossain et al.

[9]. Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization, Vol. 29. 65–72.