Generating Image Captions based on Deep Learning and Natural Language Processing

Ms. Rohini P
Asst. Professor of Department Of Computer Science and Engineering
Srinivasa Ramanujan Institute of Technology, Anantapur
rohini.cse@srit.ac.in

Bhargavi M
Department Of Computer Science & Engineering
Srinivasa Ramanujan Institute of Technology
204g1a0521@srit.ac.in

Jasmin G
Department Of Computer Science & Engineering
Srinivasa Ramanujan Institute of Technology
204g1a0542@srit.ac.in

Manjusha P
Department Of Computer Science & Engineering
Srinivasa Ramanujan Institute of Technology
204g1a0552@srit.ac.in

Anulekha Sai A
Department Of Computer Science & Engineering
Srinivasa Ramanujan Institute of Technology
214g5a0504@srit.ac.in

Abstract- Humans and computers are attempting to communicate because everything in today's society depends on systems like computers, mobile phones, etc. This is how our project is visualized. Our undertaking People with visual impairments can benefit from the creation of image captions. Computers are unable to distinguish objects, things, or activities with the same ease as humans. To recognize them, they require some training. The suggested method is used to identify activities or similar items. We offer several deep neural network-based models for creating captions for images, with a particular emphasis on CNNs (Convolutional Neural Networks) that extract characteristics from the image. Using LSTM (Long Short-Term Memory) techniques, RNNs (Recurrent Neural Networks) create captions based on the image's attributes. and examining how they affect the construction of sentences. Here, encoder-decoders are used to create a link between descriptions from natural language processing and visual information such as image features. The process of generating a caption's sequence is handled by the decoder, while the encoder extracts features. In order to determine which feature extraction and encoder model produces the best results and accuracy, we have also created captions for sample photos and compared them with one another. We also introduce Deep Voice, a text-to-speech system of production quality that uses only deep neural networks to generate captions based on visual attributes. The evaluation of our project will be conducted utilizing several machine learning methods and Python.

Keywords - CNN, RNN, LSTM, Encoder - Decoder.

1.INTRODUCTION

It is relatively easy for humans to describe the environments they are in. It is normal for a human to be able to quickly describe a vast amount of information about an image[1]. This is a fundamental human ability. The ability to identify objects and describe images is facilitated by the human brain. Artificial Intelligence introduces numerous algorithms that are based on the architecture of the brain. Here, human beings are employing these algorithms to mimic human visual world interpretation on computers. Despite significant advancements in computer vision fields including object identification, picture classification, attribute classification, and scene recognition. Allowing a computer to automatically explain an image that is forwarded to it using a language that resembles that of a person is a relatively new undertaking. Image captioning is the process of automatically producing a natural

language description for an image using a computer. It is a difficult task. Image captioning necessitates both a high level comprehension of an image's semantic contents and the ability to convey the information in a sentence that sounds human because it connects computer vision with natural language processing. Giving computers access to the visual world will open up a wide range of potential applications, including the creation of natural human-robot interactions, early childhood education, information retrieval, and support for those who are visually impaired. It is a relatively new endeavor to let a computer automatically explain an image that is forwarded to it in a language that sounds human.

The process of automatically creating a natural language description for a picture with a computer is called image captioning. It's a challenging task. Because picture captioning combines computer vision and natural language processing, it requires both a high level of knowledge of an image's semantic contents and the ability to transmit the information in a human-sounding sentence. A plethora of potential applications, such as the development of natural human-robot interactions, early childhood education, information retrieval, and assistance for individuals with visual impairments, will become possible if computers are given access to the visual world.

Convolutional neural networks (CNN) are used to extract features from the picture, while recurrent neural networks (RNN) are used to generate phrases in natural language based on the image. During the first stage, we have taken a novel technique to extracting features from a picture, which will provide us with details on even the smallest change between two comparable photos, instead of just detecting the objects present in the image. The 16 convolutional layer model VGG-16 (Visual Geometry Group) has been employed by us for object recognition.

We must use the dataset's captions to train our features in the second stage. In order to construct our phrases from the provided input photos, we employ two architectures: GRU (Gated Recurrent Unit) and LSTM (Long Short-Term Memory). We have measured the accuracy of different algorithms using the BLEU (Bilingual Evaluation Understudy) in order to estimate which architecture is best. For evaluating the caliber of translations produced by machines, BLUE offers a numerical score. Our goal is to achieve greater accuracy than previous efforts. We are utilizing the fickr30K dataset to guarantee more accuracy. Images are fed into a computer system as two-dimensional arrays, and the images are mapped to descriptive phrases or captions. The task of automatically creating image captions has received a lot of attention in the past few years.

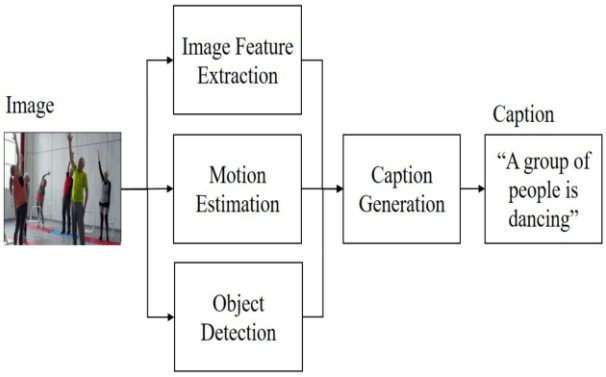


Fig1. Our model is based on a deep learning neural network that consists of a vision CNN followed by a language generating RNN. It generates complete sentences as an output….

1.1. Deep Learning

A subset of machine learning known as "deep learning" was first demonstrated by using real neurons from the brain and transforming them into artificial neural networks using learning techniques.

No matter how well-optimized, machine learning approaches begin to lose accuracy and performance as data volume increases, but deep learning performs far better in these situations.

Many domains, including image classification, speech recognition, recommendation systems, NLP (natural language processing), etc., use deep learning.

In light of all of these factors, we have decided to create our project using deep learning techniques. Features are directly learned by Deep Learning from the data. Neural networks automatically find and express pertinent patterns during training, eliminating the need for human feature creation.

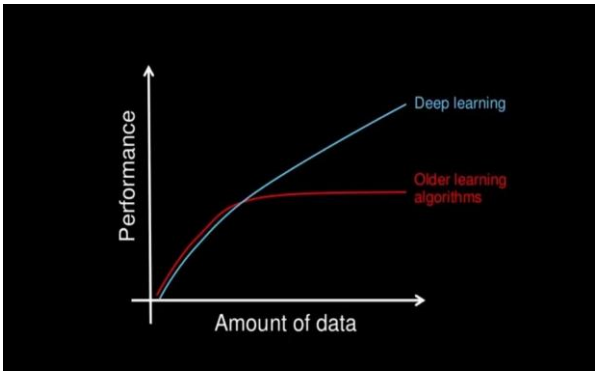


Fig2: Performance of Deep Learning

Deep learning models can adapt and generalize well to a variety of complicated datasets thanks to this capability.

Deep learning models can adapt and generalize well to a variety of complicated datasets thanks to this capability. The primary areas of attention for this project, "Generating Image Captions by using CNN and NLP," are feature extraction and classification. That being said, deep learning was chosen primarily for this reason.

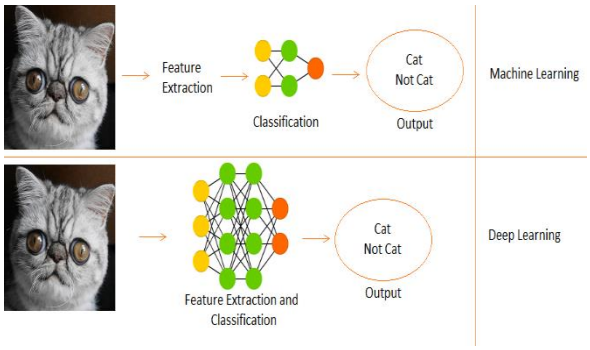


Fig3: Difference between other Machine Learning techniques and Deep Leaning

The goals of this study are to produce more accurate results for appropriate caption generation based on image attributes utilizing the flickr30K dataset. furthermore offer an audio option for the output caption that is generated. Based on the project, this paper introduces two objectives. They are as follows:

1. To create a deep learning-based image captioning system that is more accurate than the state-of-the-art in terms of captioning by effectively training the dataset using the flickr30k dataset.

2. To use text-to-speech to turn the generated image description into audio.

2.LITERATURE SURVEY

The low accuracy of the current models is what led to the creation of this research. However, since everything is digital and has displays these days, this concept will have practical implications in the modern world. Numerous investigators devised multiple methods to guarantee precision. However, some of them fall short, while some succeed to the best of their abilities.

M Sailaja, K Harika, B Sridhar, Rajan Singh. [1] In recent years, deep neural networks have enabled the captioning of images. Based on the dataset, the picture caption generator assigns a suitable title to an applied input image. The current study suggests a deep learning-based model and applies it to produce a caption for the input image. The model uses techniques such as CNN and LSTM to frame a statement connected to an image that is provided as input.This CNN model recognizes the objects in the picture, and the Long Short-Term Memory (LSTM) model not only generates the text but also a caption that fits the project.Thus, the primary goals of the suggested model are object recognition and title generation for the input photos.

C S Kanimozhiselvi, Karthika V, Kalaivani S P, Karthika S. [2] Picture captioning is the process of creating a written description for a picture. It is currently one of the more recent and pressing research issues. Different approaches to solving the issue are being introduced on a daily basis. Despite the abundance of existing options, much attention is still needed to achieve more accurate and better results. Therefore, in order to achieve better results, we thought of creating an image captioning model that combines various configurations of Long Short-Term Memory and Convolutional Neural Network architecture. For the purpose of creating the model, we combined three CNN and LSTM combinations. Three Convolutional Neural Network architectures, including Inception-v3, Xception, and ResNet50, are used to train the suggested model. These networks are used for feature extraction from the image and for producing suitable captions using Long Short Term Memory. Based on the model's accuracy, the optimal mix of three CNN and LSTM combinations is chosen. The Flicker8k dataset is used to train the model.

Chetan Amritkar, Vaishali Jabade. [3] Artificial Intelligence (AI) uses computer vision and natural language processing (NLP) to automatically synthesize an image's contents. A regenerative neuronal model is developed. It is dependent upon machine translation and vision. This strategy is employed to produce organic phrases that ultimately explain the picture. Convolutional neural networks (CNN) and recurrent neural networks (RNN) make up this paradigm. Sentence creation is done with an RNN, and feature extraction from images is done with a CNN. The model is trained so that when an input image is supplied, it produces captions that pretty much describe the image. Various datasets are used to assess the model's accuracy, smoothness, and command of language learned from picture descriptions. These tests demonstrate that the model often provides precise descriptions for the input image.

Varsha Kesavan, Vaidehi Muley, Megha Kolhekar [4]

The goal of the research is to automatically create captions by using the image's content as a source. Currently, photographs require human annotation, making the task nearly unfeasible for large commercial datasets. The Convolutional Neural Network (CNN) encoder creates a "thought vector" by utilizing the image database to extract features and nuances from the image. An RNN (Recurrent Neural Network) decoder then translates the objects and features provided by the image to produce a sequential and meaningful description of the image. In this study, we fully analyze multiple deep neural network-based picture caption generation approaches and pretrained models to determine the best efficient model with fine-tuning. To maximize the model's capacity to generate captions, they examined models that included and did not contain the "attention" concept. For a more accurate comparison, every model is trained using the same dataset.

T J Buschman, E K Miller. [5] In the human visual system,bottom-up signals linked to unexpected, unusual, or salient stimuli can automatically focus attention, as can top-down signals dictated by the current task (e.g., looking for something). In this work, they use comparable language to designate as "top-down" attention mechanisms those that are driven by nonvisual or task-specific context, and a "bottom-up" attention mechanisms that are solely visual feed-forward. The majority of traditional visual attention processes utilized in VQA and picture captioning are top-down in nature. Usually taught to selectively attend to the output of one or more layers of a convolutional neural net (CNN), these methods take as context a representation of a partially-completed caption output, or a question related to the image. Nevertheless, this method pays minimal attention to the process of selecting the visual parts that require attention. It is a time taking process to pre-process the image.

Farhadi et al. [6] The three primary categories used in this paper are picture captioning techniques are covered in this section i.e., template-based image captioning, retrieval-based image captioning, and novel caption creation. In template-based approaches, captions are generated using preset templates that contain blank spaces. These systems fill in the blanks in the templates after first identifying the various objects, actions, and characteristics. To generate image captions, for instance, complete the template slots with three distinct scene pieces.

Lakshmi Narasimhan Srinivasan, Dinesh Sreekanthan, A L Amutha. [7] The keras framework's TensorFlow backend has been utilized in this study's model evaluation. Utilizing assessment measures that were appropriate for the problem's nature allowed for an understanding of how The model has made correct predictions. This paper presents the results of mathematical computations performed on the confusion matrix.

D Elliott, F Keller [8]. The main difficulties in this research include identifying the objects in an image and their characteristics, which are challenging computer vision problems, as well as figuring out how the objects interact and what relationships exist between them. Automatic image description is not without its difficulties. To improve the model's performance, the authors trained it over several layers (or levels) using CNN.

C Amritkar and V Jabade [9]. In this paper, the model is trained so that it can produce captions that almost perfectly describe an input image when it receives one. Several datasets are used to test the model's correctness as well as the language model's smoothness or command after learning from picture descriptions. These tests demonstrate that the model often provides precise descriptions for the input image.

V Kesavan, V Muley, M Kolhekar [10]. The goal of the article is to use image content to generate captions automatically. This work presents a comprehensive analysis of various deep neural network-based image techniques for creating captions and pre-trained models to identify the most effective model and fine- tune it. To improve the caption, they examined models with and without attention ideas. Producing capacity of the model. For a more accurate comparison, every model is trained using the same dataset.

Kulkarni et al. [11] Before filling in the blanks, Kulkarni uses a Conditional Random Field (CRF) to identify the objects, attributes, and prepositions. Though the templates are established, template-based techniques can produce grammatically correct captions but not variable-length captions. They go over the three primary categories of current image in this part. Approaches for captioning images: retrieval-based captioning, template-based captioning, and creative caption creation. For the purpose of generating captions, template-based solutions use predetermined templates with blank spaces. Within these systems, the many items, actions, and after

identifying the qualities, the gaps in the templates are filled.

N K Kumar, D Vigneswari, A Mohan, K Laxman, J Yuvraj [12]. This work uses deep learning to discover, recognize, and produce meaningful captions for a given image. For object identification, recognition, and caption generation, Regional Object Detector (ROD) is utilized. The suggested approach leverages deep learning to enhance the current image caption generation system even more. Python is used to conduct experiments on the Flickr 8k dataset in order to illustrate the suggested approach. We are using the best and large dataset i.e., Flickr 30k.

3.PROPOSED SYSTEM

The project's suggested system is an advanced and adaptable picture captioning solution that combines natural language processing (NLP) with deep learning techniques to produce precise, linguistically coherent, and contextually relevant captions for images with accompanying audio. CNN is given an image to evaluate and create a feature vector from. This feature vector enhances the user's perception of the image by providing an auditory context that corresponds with the visual content. It is used as input for sigmoid functions and RELU functions in the GRU and LSTM. It also provides image descriptions and is associated with pertinent sounds for the image.

3.1 Architecture:

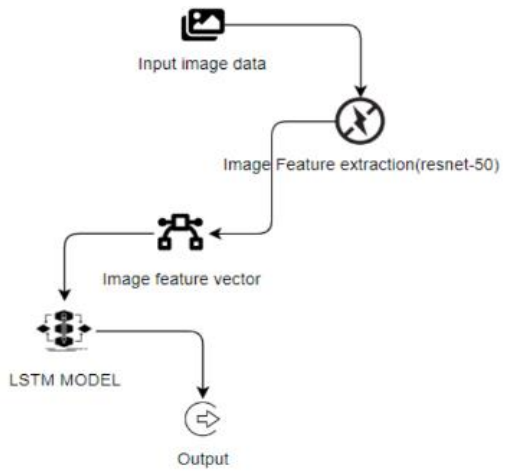


Fig3: Architecture

3.2 System :

3.2.1 Create Dataset:

To assess the performance of the model, the dataset comprising text and image data of the intended objects to be captioned is divided into training and testing datasets.

3.2.2 Pre-processing:

Prior to being input into a machine learning model, images must be enhanced and prepared. To train our model, we resize and reshape the photos into the proper format.

3.2.3 Training:

We train our model utilizing the CNN and LSTM algorithms by using the pre-processed training dataset.

3.3 User:

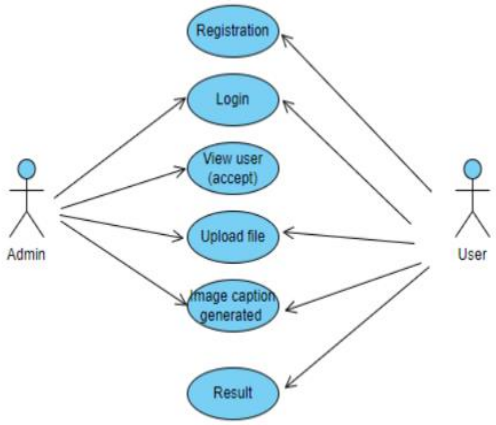


Fig4. Use Case Diagram

3.3.1 Register:

The user must register. The information is kept in a database, and the administrator will review it and grant access if the administrator determines that the registered information is accurate.

3.3.2 Admin Login:

The administrator logs in and examines the list of users who have registered. Only the user will be able to log in when he has approved the user data. The administrator has complete control over who may read, edit, and manage the data.

3.3.3 Login:

Users can use this technique to authenticate themselves with the system and access the application by entering their credentials, which consist of their username and

password.

3.3.4 Upload image:

The user must choose an image from the dataset and upload it into the program. The image must have a caption.

3.3.5 Prediction:

It produces a caption as an output, and the image caption that we have given to it will be displayed as a result of our model's operation.

3.3.6 Logout:

The user has the option to exit the application when the result has been generated.

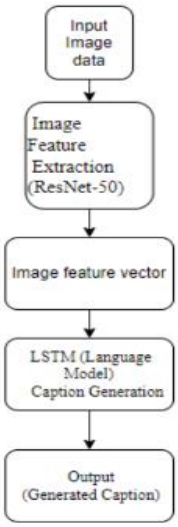


Fig5. Block Diagram

4.IMPLEMENTATION

4.1 CNN:

Over the past 20 years, Deep Learning has shown to be a very powerful technology due to its ability to handle large amounts of data. Convolutional neural networks, also referred to as CNNs or ConvNets, are among of the most popular deep neural networks in deep learning, especially for applications in computer vision. The type of deep neural network utilized in deep learning that is most commonly used to evaluate visual data is called a convolutional neural network (CNN). It employs a special technique called convolution. As it is currently understood, convolution is a mathematical operation on two functions that results in a third function that characterizes the transformation of one's shape by another.

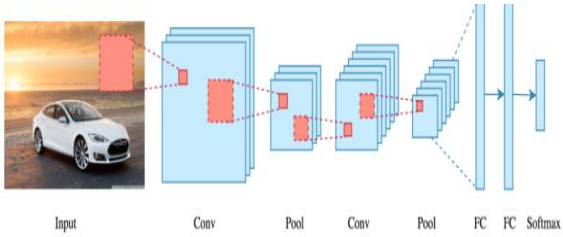


Fig6. Working of CNN

Convolutional, pooling, and fully-connected (FC) layers are the three types of layers that make up a CNN. These layers are stacked to construct a CNN architecture. Beyond these three layers, two more important factors are the dropout layer and the activation function. The RGB (Red, Green, and Blue) color space, which extends from 0 to 255, can be used to extract and identify various visual aspects for analysis (a process known as feature extraction) using a convolution tool. The network used for feature extraction is composed of multiple pairs of convolutional or pooling layers.

4.2 ResNet 50:

A kind of deep neural network architecture called ResNet (Residual Network) was created to solve the vanishing gradient issue that arises when deep convolutional neural networks (CNNs) are being trained. ResNet gives the network the ability to learn residual functions by introducing skip connections, also referred to as residual connections. The model can bypass specific levels thanks to these skip connections, which send the input straight to the output of deeper layers. This facilitates the training of extremely deep networks by reducing the impact of the vanishing gradient issue.

ResNet can be a key component in feature extraction from photos within an image caption generator. In order to extract useful features from the input photos, the encoder portion of an image captioning model usually makes use of a CNN that has already been trained, such as ResNet.

The goal is to extract high-level features from photos by using the information that the pre-trained ResNet model has acquired on a sizable dataset (such as ImageNet).

- Images are used to extract features using the ResNet model. For image classification tasks, the model is usually pre-trained on a large

dataset (e.g., ImageNet). Hierarchical and abstract aspects in photos are captured by the weights that were learned during pre-training.

- The pre-trained ResNet model is used to extract features from intermediate layers given an input image. The image's high-level visual information is represented by the features.
- The decoder component of the picture captioning model receives the features that were extracted from the image. Based on the input attributes, the decoder—which is frequently implemented as a transformer or recurrent neural network (RNN)—creates a textual description of the image.

4.3 RNN:

A type of neural network called a recurrent neural network (RNN) uses the output from the preceding step as the input for the current step. All of the inputs and outputs of conventional neural networks are independent of one another. However, in situations when it is necessary to guess the following word in a sentence, the preceding words are necessary, hence it is necessary to retain the preceding words. Thus, RNN was created, and it used a Hidden Layer to tackle this problem. The Hidden state of an RNN, which retains some information about a sequence, is its primary and most significant feature. Because the state retains memory of the prior input to the network, it is also known as Memory State.

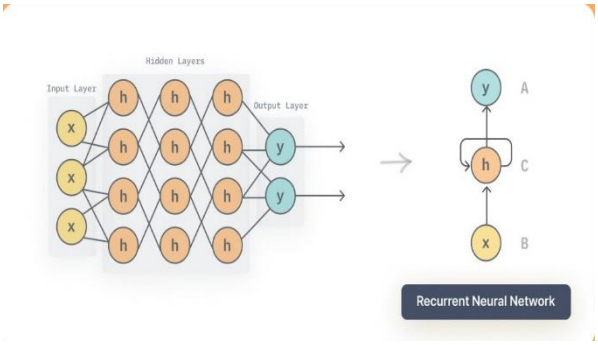


Fig7: Single RNN

It does the same task on all inputs or hidden layers to produce the output, using the same settings for each input. Unlike other neural networks, this lowers the parameter complexity. Although it isn't referred to as a "recurrent neuron," a recurrent unit is the basic processing unit of a recurrent neural network (RNN). Because of its special capacity to preserve a hidden state, this unit enables the network to recognize sequential relationships by processing and remembering prior inputs. The input and output architecture of RNNs is identical to those of other deep neural architectures. The way information moves from input to output, however, varies.

Utilizing identical parameters for every input, it performs the same operation on all inputs or hidden layers to generate the output. Its decreased parameter complexity sets it apart from other neural networks. Despite not being called a "recurrent neuron," a recurrent unit is the fundamental processing unit of a recurrent neural network (RNN). This unit's unique ability to maintain a concealed state allows the network to process and retain previous inputs, which in turn helps the network understand sequential relationships. Like other deep neural architectures, RNNs have the same input and output architecture. Yet, there are differences in the way data is transferred from input to output.

The formula for calculating the current state:

$h_t = f(h_{t-1}, x_t)$

Where, h_t -> current state

h_{t-1} -> previous state

x_t -> input state

Formula for applying Activation function(tanh):

$h = \tanh(W_{hh}h_{t-1} + W_{xh}x_t)$

Where, W_{hh} -> weight at recurrent neuron

W_{xh} -> weight at input neuron

The formula for calculating output:

$Y_t = W_{hy}h_t$

Where, Y_t -> output

W_{hy} -> weight at output layer

4.4 LSTM:

The Long Short-Term Memory, or LSTM, is an enhanced RNN. For sequence prediction tasks, LSTM performs remarkably well in capturing long-term dependencies.. The RNN method has certain drawbacks, which we address by introducing the LSTM algorithm.

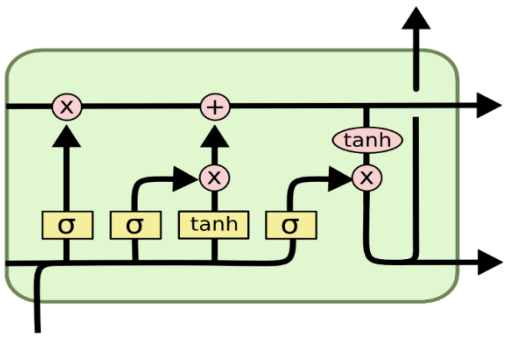


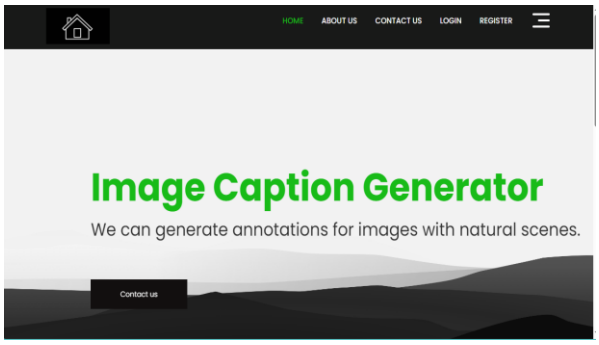
Fig8: Architecture of LSTM

- It can be challenging for a network to learn long-term dependencies in a standard RNN since it only has one hidden state that is retained over time. In contrast, LSTMs solve this issue by introducing memory cells, which are containers that can store information for a longer amount of time.
  Three gates govern a memory cell:i. Input gate
  ii. Forget gate
  iii. Output gate
- These gates determine which data should be input into, taken out of, and output from the memory cell.
- Unlike RNNs, which lack memory units, LSTMs have a unique memory unit that enables them to recognize long-term dependencies in sequential data.
- While RNN is also meant to process sequential data, its memory capacity is constrained. In contrast, LSTM is well suited for handling sequential data.
- Compared to RNN, the training process of LSTM is slower because of its increased complexity. Because of its more straightforward architecture, RNNs train a little bit faster.
- Long sequences are more effective for LSTM, but RNN finds it difficult to store information.

5. RESULTS AND ANALYSIS:

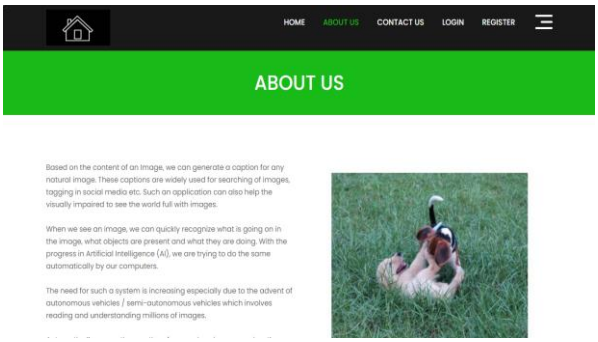
5.1 Home page:

This is the home page where we land after clicking on the link.



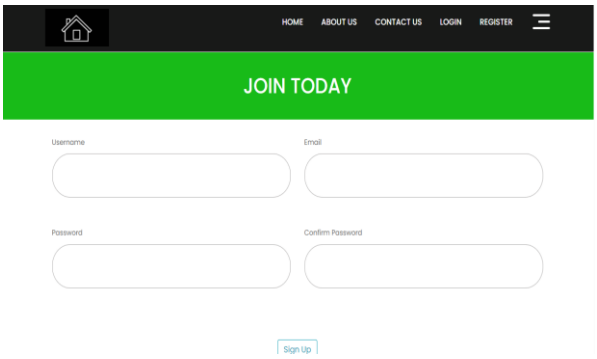
5.2 About Page:

Here we have a slight description about the project.



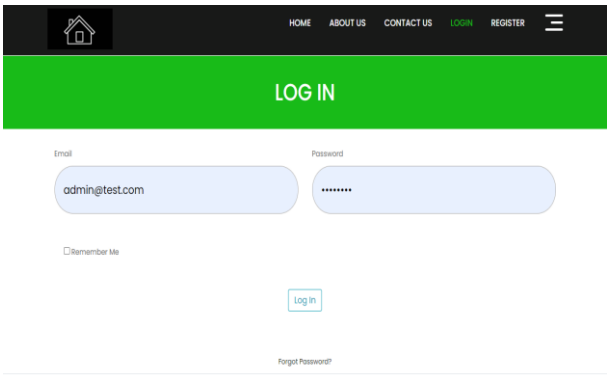
5.3 Register Page:

Here User registers themselves.



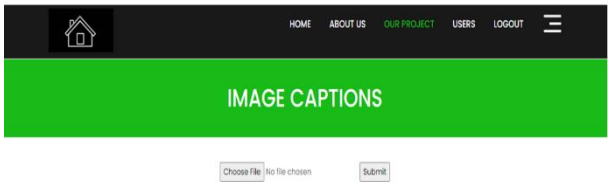
5.4 Login Page:

Here user logs in with the credentials they registered with.



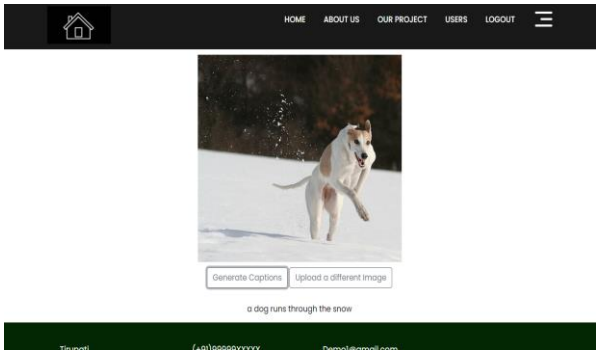
5.5 Upload Page:

Here user uploads the image and caption is generated.



5.6 Results:

Here we get the results.



6. CONCLUSION AND FUTURE WORK

The issue of creating meaningful captions for images has been found to be powerfully and effectively solved by the image caption generator that combines Long Short-Term Memory (LSTM) networks with Convolutional Neural Networks (CNNs). Using CNN layers to extract relevant features and capture spatial information, the CNN-LSTM model showed how to sequence and generate coherent and contextually appropriate captions by efficiently utilizing LSTM layers. Visual perception and sequential data processing work together to address the problems of picture understanding and natural language synthesis through the integration of these two architectures. In addition to demonstrating the promise of deep learning for multimodal tasks, this work emphasizes the need of merging specialized neural networks to produce better results in challenging tasks like image captioning.

The CNN and LSTM image caption generator can be improved and expanded upon in a number of ways in future work. Firstly, a better representation of the intricate relationships between textual and visual data may be achieved by the model by exploring more sophisticated architectures such as attention processes, transformer models, or language models that have previously undergone training, such as BERT.

Furthermore, adding a larger and more varied dataset to the training set can improve the model's generalization and make it capable of accurately describing a wider variety of images. It could also be beneficial to fine-tune the model for particular domains or activities, enabling the generator to specialize in fields like satellite imagery or medical imaging. Additionally, researching methods to improve the interpretability and controllability of the model may help to improve comprehension and guidance of the captioning process. Last but not least, putting the model to use in actual applications and getting user input would shed light on its applicability in the real world and point out possible improvements.

7. REFERENCES

- [1]. M Sailaja, K Harika, B Sridhar. Rajan Singh, “Image Caption Generator using Deep Learning”, 2022 International Conference on Advancements in Smart, Secure and Intelligent Computing (ASSIC).
- [2]. C S Kanimozhiselvi, Karthika V, Kalavani S P, Krithika S, “Image Captioning Using Deep Learning”, 2022 International Conference on Computer Communication and Informatics (ICCCI).
- [3]. Chetan Amritkar, Vaishali Jabade, “Image Caption Generation Using Deep Learning Technique”, 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA).
- [4]. Varsha Kesavan, Vaidehi Muley, Megha Kolhekar, “Deep Learning based Automatic Image Caption Generation”, 2019 Global Conference for Advancement in Technology (GCAT).
- [5]. T. J. Buschman and E. K. Miller. “Top-down versus bottom-up control of attention in the prefrontal and posterior parietal cortices. Science”, 315(5820):1860–1862, 2007.
- [6]. William Fedus, Ian Goodfellow, Andrew M Dai. Maskgan, “Better text generation”, 1801.07736, 47, 2018.

- [7]. Lakshmi narasimhan Srinivasan, Dinesh Sreekanthan and A.L Amutha, “Image captioning - A Deep Learning Approach”, International Journal of Applied Engineering Research ISSN 0973-4562 Volume 13, Number 9 (2018) pp.

- [8]. D. Elliott, F. Keller, “Image Description using Visual Dependency Representations”, Conference on Empirical Methods in Natural Language Processing.

- [9]. C. Amritkar, V. Jabade, “Image Caption Generation using Deep Learning Technique”, IEEE Access, 2018.

- [10]. V. Kesavan, V. Muley and M. Kolhekar, “Deep Learning based Automatic Image Caption Generation”, IEEE Access, 2019.

- [11]. Girish Kulkarni, Visruth Premraj, Sagnik Dhar, Siming Li, Yejin Choi, Alexander C Berg, and Tamara L Berg. “Baby talk: Understanding and generating image descriptions”, IEEE Transactions on Pattern Analysis and Machine Intelligence, 35:2891–2903, June 2013.

- [12]. N K Kumar, D Vigneswari, A Mohan, K Laxman, J Yuvaraj, “Detection and Recognition of Objects in Image Caption Generator System: A Deep Learning Approach”, IEEE – 2019.