



ISSN: 2350-0328

**International Journal of Advanced Research in Science,  
Engineering and Technology**

**Vol. 12, Issue 4, April 2025**

# Visual Image Caption Generator Using Deep Learning

**Azmeera NavyaSri, Bakthula KalyaniSai, Begary Deepthi, Dr Y V S S Pragathi**

Student, Department of Computer Science and Engineering, Stanley College of Engineering and Technology for Women, Hyderabad, Telangana, India

Student, Department of Computer Science and Engineering, Stanley College of Engineering and Technology for Women, Hyderabad, Telangana, India

Student, Department of Computer Science and Engineering, Stanley College of Engineering and Technology for Women, Hyderabad, Telangana, India

Professor, Department of Computer Science and Engineering, Stanley College of Engineering and Technology for Women, Hyderabad, Telangana, India

**ABSTRACT:** This project aims to develop a deep learning system that automatically generates meaningful descriptions for images. With advancements in deep learning, combining computer vision and natural language processing has become a key research area. The system describes the context of a photograph in English using one or more sentences. It interprets visual data by analyzing object states, properties, and relationships. A CNN-LSTM architecture is employed, where CNN extracts features from greyscale images and LSTM generates captions. The 2017 COCO Dataset is used for training and evaluation. The model identifies objects and sends text notifications based on image content. To aid the visually impaired, the system can be extended to provide spoken messages. This generative model aims to outperform traditional methods and human baselines in image captioning.

**KEYWORDS:** Deep Learning, LSTM, CNN, NLP, Computer Vision.

## I. INTRODUCTION

Image captioning leverages deep learning to bridge computer vision and natural language processing, enabling machines to generate descriptive text for images. This is achieved through an encoder-decoder architecture, where a Convolutional Neural Network (CNN) encodes visual features, and a Long Short-Term Memory (LSTM) network decodes these features into coherent sentences. Such systems are instrumental in aiding visually impaired individuals by converting visual content into speech, enhancing their interaction with the environment. Beyond accessibility, image captioning finds applications in various domains, including social media, e-commerce, surveillance, and autonomous vehicles, by providing contextual understanding of visual data. The integration of CNN and LSTM models facilitates the generation of accurate and contextually relevant captions, contributing to advancements in artificial intelligence and machine learning.

## II. RELATED WORKS

### Table 1: Literature Survey

**Miss. Bibi Zinab Dongarkar and Miss. Simran Sarang et.al., (2023) [1]** This study develops an image caption generator using a CNN-LSTM architecture, where CNNs extract visual features from images, and LSTMs generate descriptive captions. Trained on the COCO dataset, the model faces challenges like limited dataset size, poor generalization, and difficulty in handling complex scenes with multiple objects.

**Shourya Tyagi, Olukayode A. Oki, Vineet Verma, and Swati Gupta et.al., (2024) [2]** The ICTGAN model integrates Vision Transformers (ViTs), Generative Adversarial Networks (GANs), and LSTMs for enhanced image captioning. ViTs capture detailed visual features, while GANs improve caption quality. Despite its advancements, the model faces challenges such as complexity and resource demands.



ISSN: 2350-0328

# International Journal of Advanced Research in Science, Engineering and Technology

Vol. 12, Issue 4, April 2025

**Praveen Kumar Tripathi and Dr. Shambhu Bharadwaj et.al., (2024) [3]** This survey explores hybrid deep learning techniques in image captioning, focusing on CNNs for feature extraction and RNNs (LSTM, Bi-LSTM, GRU) for caption generation. The paper discusses challenges like model complexity, overfitting, and dataset dependency.

**Mohammed Inayathulla and Karthikeyan C et.al., (2024) [4]** The authors propose a CNN-LSTM approach for image caption generation in video summarization, using DenseNet201 for feature extraction and GloVe for embedding. The model's performance is hindered by limited training data and poor scalability for diverse video datasets.

**Vijay A. Sangolgi, Mithun B. Patil, Shubham S. Vidap, Satyam S. Doijode, Swayam Y. Mulmane, and Aditya S. Vadaje et.al., (2024) [5]** MVBICG is a multilingual, real-time captioning system that integrates CNNs and LSTMs with Google Translate and Text-to-Speech for accessibility. The model faces issues with dataset size, translation biases, and scalability.

**Varsha Kesavan and Vaidehi Mely et.al., (2019) [6]** This deep learning model uses CNNs for feature extraction and LSTMs for caption generation. It highlights the challenges of complex scenes and ambiguous contexts in caption generation.

**Dr. Aziz Makandar and Keerti Suvarnakhandi et.al., (2022) [7]** The authors develop a CNN-LSTM model using VGG16 for feature extraction and LSTM for generating captions. It is trained on the Flickr8k dataset but is limited by a small dataset, lack of multilingual support, and no real-time captioning.

**Chetan Amritkar and Vaishali Jabade et.al., (2019) [8]** The model employs CNNs for visual feature extraction and LSTMs for generating captions. It struggles with complex scenes, ambiguous contexts, and the need for large, diverse datasets to improve generalization.

**Grishma Sharma, Priyanka Kalena, Aromal Nair, Nishi Malde, and Saurabh Parkar et.al., (2019) [9]** This system uses VGG16 for image feature extraction and both LSTM and GRU architectures in a Merge framework for caption generation. It faces challenges like dataset limitations and caption inaccuracies.

**Prof. A. S. Narote, Kunal Vispute, Harshit Himanshu, Rajas Bhagatkar, and Sneha Jadhav et.al., (2023) [10]** The authors propose a CNN-LSTM-based image captioning system with a web interface for real-time caption generation. Trained on the Flickr8k dataset, it performs well but faces issues with complex scenes, dataset size, and fluency of captions.

S. No	YEAR	TITLE	AUTHOR	PROPOSED WORK	METHODOLOGIES	LIMITATIONS
01	2023	Image Caption Generator using Deep Learning	Miss. Bibi Zinat Dongarkar, Miss. Simran Sarang	Automatic image caption generation using CNN and LSTM	CNN for feature extraction, LSTM for caption generation, COCO dataset	Limited dataset, lack of contextual awareness, bias in captions, high computational cost
02	2024	Novel Advance Image Caption Generation Utilizing Vision Transformer and GANs	Shourya Tyagi, Olukayode A. Oluwalanlajo, Vineet Verma, Swati Gupta	Hybrid model (ICTGAN) using ViTs for feature extraction, GAN for caption real-time generation and LSTMs for language generation	Vision Transformer, LSTM, GANs, trained on MS COCO, BLEU, ROUGE-L, CIDEr	Object hallucination, high resource demand, context understanding issues, data dependency
03	2024	A Comprehensive Survey on Image Description Generation Techniques	Praveen Kumar Tripathi, Dr. Shambhu Bharadwaj	Hybrid model (ICTGAN) using ViTs for feature extraction, GAN for caption real-time generation and LSTMs for language generation	CNN (ResNet, VGG, Inception), LSTM, Bi-LSTM, GRU, BLEU, METEOR, CIDEr	High complexity, overfitting risk, semantic mismatch, language issues, RNN limitations



04	2024	Image Caption Generation using Deep Learning for Video Summarization Applications	Mohammed Inayathulla, Karthikeyan C	CNN-LSTM framework for v summarization	DenseNet201 for image encoding, LSTM decoder trained on Flickr8k, BLEU score	Dataset size (Flickr8k), multilingual bias, scalability issues, requires internet access
05	2024	Enhancing Cross Linguistic Image Caption Generation with Indian Multilingual Voice Interfaces	Vijay A Sangolgi, Mithun B Patil, Shubham S Vidya, Satyam S Doijode, Swayam Y Mulmane, Aditya Vadaje	MVBICG system for multilingual, voice-based image captions	CNN for image feature extraction, RNN/LSTM for caption generation, Google Translate API, gTTS for text-to-speech	Dataset size (Flickr8k), multilingual bias, scalability issues, requires internet access
06	2019	Deep Learning based Automatic Image Caption Generation	Varsha Kesavan, Vaidehi Mely	Using CNN and LSTM to generate meaningful captions	CNN for feature extraction, LSTM for caption generation	Challenges in generating captions for complex images, ambiguous content, grammatical correctness
07	2022	Image Caption Generator Using CNN-LSTM	Dr. Aziz Mankar, Keerti Suvarnakhandi	CNN for image feature extraction and LSTM for caption generation	CNN as encoder, LSTM decoder, trained on Flickr8k	Limited to English, lacks multilingual support, potential overfitting
08	2019	Image Caption Generation using Deep Learning Technique	Chetan Amritkar, Vaishali Jabade	Deep learning approach for generating descriptive captions	CNNs for feature extraction, RNNs/LSTM for caption generation	Challenges with complex images, ambiguous content, grammatical correctness
09	2019	Visual Image Caption Generator Using Deep Learning	Grishma Sharma, Priyanka Kalena, Aromal Nair, Nilesh Malde, Saurabh Parkar	Image captioning system using CNN and RNN-based decoders	VGG16 for feature extraction, LSTM/GRU for decoding, trained on Flickr8k	Small dataset, may generate incorrect captions, lacks diversity in vocabulary
10	2023	Image Caption Generator Using Deep Learning Approach	Prof. A. S. Narasimhan, Kunal Vispute, Harshit Himanshu, Rajas Bhagatkar, Sneha Jadhav	Image caption generation with CNN and LSTM	CNN for feature extraction, LSTM for decoding, combines computer vision and NLP	Limited by Flickr8k dataset, challenge with complex scenes and multiple object relationships

### III. PROPOSED METHODOLOGY

The primary goal of this research is to create an image captioning system that combines Long Short-Term Memory (LSTM) networks for sequential language creation with Convolutional Neural Networks (CNN) for visual feature extraction. The system is specifically made for accessibility applications, such as helping blind and visually impaired people with activities, and it strives to provide meaningful and contextually accurate captions for images, outperforming models.

**Preprocessing and Dataset:** Prior to processing, the input images are scaled and transformed to LSTM to guarantee consistency and lower noise. Training and validation are done using the popular COCO Dataset 2017, which has a variety of annotated images. To get them ready for training, the annotations including captions are tokenised and incorporated into a lexicon.

**Feature Extraction Using CNN:** Feature extraction High-level feature representations are extracted from the input photos in this research using a pretrained CNN model. These feature vectors, which come from the CNN fully connected layer, represent crucial visual semantics such object characteristics, spatial relationships, and image context.

**LSTM Caption Generation:** To create captions in a sequential fashion, the CNN feature vectors are fed into an LSTM network. The procedure begins with a unique token, and the LSTM uses the image context and the previously created word to predict each successive word. Words are represented as dense vectors using an embedding layer, and the most likely word is chosen at each timestep using a Softmax layer.

**Model Evaluation:** Cross-entropy loss is used to train the CNN-LSTM model from beginning to end, optimising both the language generation and visual aspects at the same time time. Standard criteria including BLEU, METEOR, and CIDEr are used to assess the system's performance and guarantee high-quality caption output.

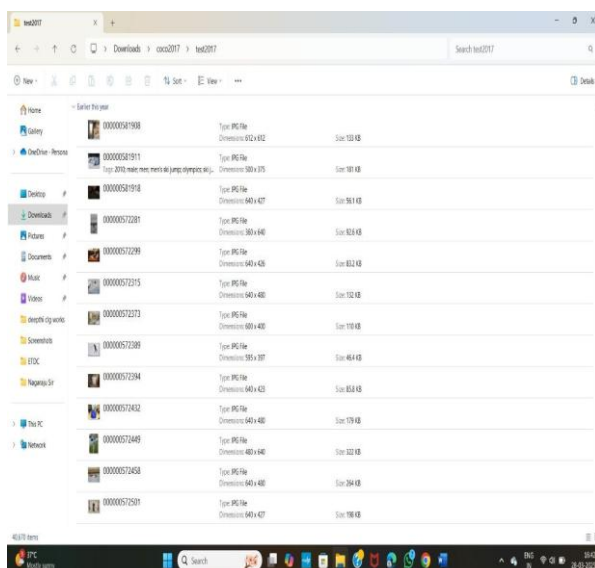


Fig: Dataset of test (COCO)

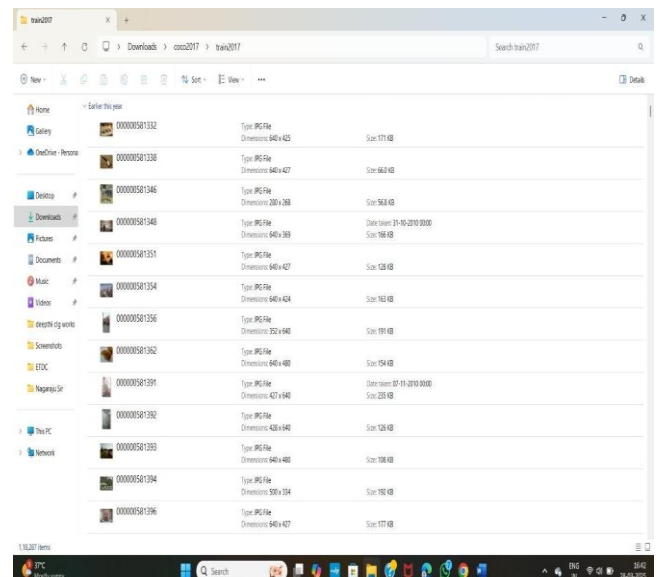


Fig : Dataset for Train(COCO)

## A. CONVOLUTIONAL NEURAL NETWORK(CNN)

Convolutional Neural Networks (CNN) are deep learning models built to process structured data like images by extracting key features through scanning from various directions. They handle image transformations such as resizing, rotation, and alterations while assigning values to different components for accurate interpretation. Unlike traditional models, CNN require minimal pre-processing and can automatically learn features. Their architecture is inspired by the human visual cortex, with neurons responding to specific regions of the image. A traditional neural network, where each neuron connects to every neuron in the next layer, becomes inefficient for processing large images due to excessive parameters and overfitting. CNN solve this by adopting a 3D architecture, where neurons focus on small sections of the image. Each group of neurons specializes in identifying specific features like a nose or an ear. This approach ensures better recognition by emphasizing important parts of the image. The final output reflects the probability of the image belonging to a certain class.

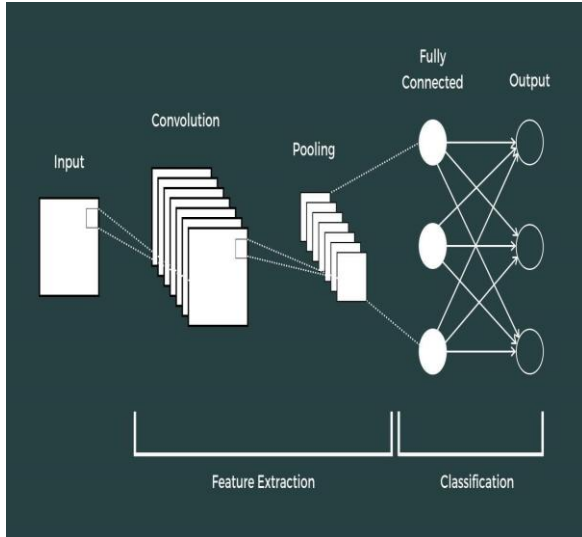


Fig : CNN Architecture

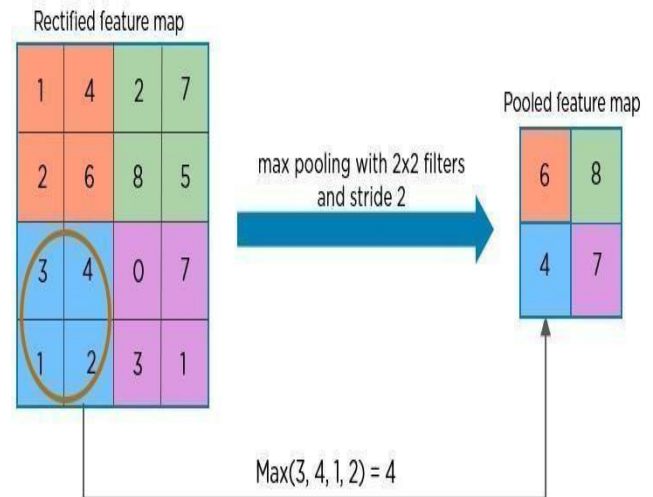


Fig : Working of CNN

## B. Long Short Term Memory

One kind of recurrent neural network that can learn order dependence in sequence prediction tasks is called an LSTM network. In complicated problem fields like speech recognition and machine translation, among others, this behaviour is necessary. LSTM are a complex area of deep learning. In complicated problem fields like speech recognition and machine translation, among others, this behaviour is necessary. LSTM are a complex place of deep learning.

Advantage of LSTM is:

- Provides us with a large range of parameters such as learning rates, and input and output biases.

## C. CNN-LSTM Model

The CNN LSTM shape consists of using Convolutional Neural Network (CNN) layers For characteristic extraction on enter facts blended with LSTMs to help series prediction. CNN-LSTM have been evolved for visible time collection prediction issues and the application of generating textual descriptions from sequence of image (e.g., videos) Specifically, the problem of creating a textual description of an activity shown in a series of pictures is known as "activity recognition"

- Image Description: Producing an explanation in prose for a single image. Creating a textual explanation of a series of images is known as "video description generation. We shall use the more general term "CNN LSTM," but the original designation for this architecture was a Long-term Convolutional Network (LCN) model.
- To extract features from the image, CNN is utilised. Xception, a pre-trained model, will be employed.
- LSTM will assist in creating an image description by utilising the CNN data.

## D. FLOW CHART

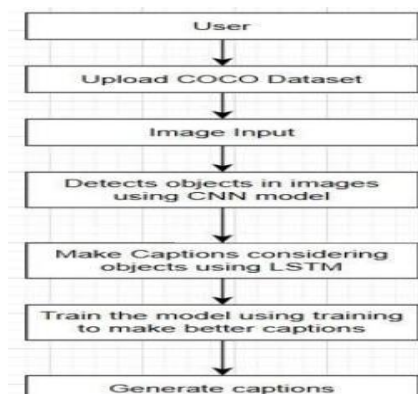


Fig: Flow Chart



## E. SYSTEM ARCHITECTURE

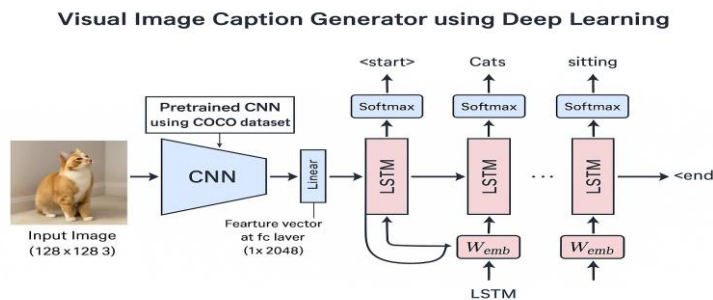


Fig: Proposed System Architecture

Fig represents the architecture for Visual Image Caption Generation integrates Convolutional Neural Networks (CNN) for characteristic extraction and Long Short-Term Memory (LSTM) networks for sequence generation. This hybrid model is designed to process an input image and generate a meaningful textual description that captures the objects, their attributes, and their relationships in the scene.

The proposed methodology integrates Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks to generate accurate image captions. Initially, images are resized to  $224 \times 224 \times 3$  and processed through a pretrained CNN (e.g., Xception) to extract high-level visual features, resulting in a  $1 \times 1 \times 2048$  feature vector. This vector encapsulates semantic information like object presence and spatial relationships. Subsequently, a linear layer transforms this feature vector to align with the LSTM input requirements. The LSTM network then generates captions by converting words into embedding vectors and predicting subsequent words in the sequence, utilizing a Softmax layer at each step to determine the most probable word. This iterative process continues until an end token signifies the completion of the caption. The final output is a coherent sentence describing the image content, such as "A cat sitting in an open area." This architecture effectively bridges computer vision and natural language processing, enhancing applications like automated image annotation, assistance for visually impaired individuals, and improved human-computer interactions.

## STEPS FOR IMPLEMENTATION:

1. The user uploads an image for caption generation.
2. The image is processed through a CNN, which scans and extracts key features using layers like Convolutional, Pooling, and Fully Connected.
3. The extracted features are passed to an LSTM, which predicts and generates a descriptive sentence for the image.

## IV. RESULTS AND DISCUSSION

**1. Accuracy:** The total percentage of accurate predictions the model makes is known as accuracy. The percentage of words (or n-grams) in the generated caption that precisely match the ground truth caption can be used in the context of picture captioning.

$$\text{Accuracy} = \frac{TP + TN + FP + FN}{TP + TN}$$

**2. Precision:** Precision quantifies the proportion of pertinent and accurate words in the generated caption (i.e., those that match the ground truth caption).  $\text{Precision} = \frac{TP + FP}{TP}$

Where, True Positives (TP): The number of words in the generated caption that match words in the ground truth caption.

False Positives (FP): The number of words in the generated caption that do not match any word in the ground truth caption.

**3.Recall:** Recall measures how many of the relevant words in the ground truth caption were captured by the generated caption. **Recall =  $\frac{TP}{TP+FN}$**

Where, True Positives (TP): The number of words in the generated caption that match words in the ground truth caption.

False Negatives (FN): The number of words in the ground truth caption that are missing in the generated caption.

**4.F1-Score:** The F1 Score is the harmonic mean of Precision and Recall. It is a balance between Precision and Recall, useful when you need to consider both false positives and false negatives equally.

**F1-Score =  $\frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$**

## V. OUTPUT DESIGN

a blue butterfly on a white background  
a blue butterfly with a transparent blue body  
the dark blue moro butterfly has been named for its brilliant plumage the dark blue  
moro butterfly has been named for its wonderful black and blue  
blue butterfly of a butterfly, in a cut out position isolated on a white background  
image  
blue butterfly isolated on a white background



**Fig : Output**

Fig describes On a crisp white background, the image displays a detailed digital illustration of a butterfly—more precisely, a blue morpho. The most remarkable aspect of the butterfly is its wings, which have a vivid, nearly iridescent blue tint that changes to black borders with white dots. Clearly visible are the elaborate vein patterns on the wings, which enhance the illustration's realism. The dark body of the butterfly stands out sharply against the vivid blue wings. As is typical of butterflies, the antennae are slender and fragile. The butterfly's beauty and complex wing patterns are highlighted by the composition's overall clarity and concentration. The picture, which highlights the butterfly's distinctive colour and shape, seems to have been produced for artistic or scientific reasons.

## VI. CONCLUSION

By bridging the gap between natural language synthesis and visual perception, the CNN LSTM model for automatic picture captioning represents a significant achievement in artificial intelligence. With the help of Long Short-Term



ISSN: 2350-0328

# International Journal of Advanced Research in Science, Engineering and Technology

Vol. 12, Issue 4, April 2025

Memory (LSTM) networks to produce coherent and contextually appropriate words and Convolutional Neural Networks (CNN) to extract meaningful characteristics from images, this model shows a strong capacity to describe visual material. Although the model demonstrates encouraging outcomes in converting intricate visual data into descriptions that are comprehensible to humans, difficulties still exist in reaching perfect precision and contextual correctness. It is imperative that training methods and model architectures be further improved in order to increase the model's capacity to produce correct and nuanced captions, which will open up a variety of uses for automated picture analysis and retrieval as well as accessibility enhancements.

## REFERENCES

- [1] Miss.Bibi Zainab Dongarkar, Miss.Heena Dongarkar, Miss.Simran Sarang, "Image Caption Generator Using Deep Learning", International Journal of Creative Research Thoughts (IJCRT) , Volume 11, Issue 10 October 2023 ISSN: 2320-2882
- [2] Shourya Tyagi ,Swati Gupta , " Novel Advance Image Caption Generation Utilizing Vision Transformer and Generative Adversarial Networks", Computers to 13120305, Issued in 2024
- [3] Praveen Kumar Tripathi, Dr. Shambhu Bharadwaj, "A Comprehensive Survey on Image Description Generation Techniques ", International Journal of Creative Research Thoughts (IJCRT) , Volume 12, Issue 7 July 2024 | ISSN: 2320-2882
- [4] Mohammed Inayathulla, Karthikeyan, " Image Caption Generation using Deep Learning For Video Summarization Applications", (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 15, No. 1, 2024
- [5] Vijay A Sangolgi, "Enhancing Cross -Linguistic Image Caption Generation with Indian Multilingual Voice Interfaces using Deep Learning Techniques ", Elsevier B.V , 233, issued in 2024
- [6] varsha kesavan, vaidehi mely , Megha Kolhekar , "Deep Learning based Automatic Image Caption Generation", Institute of Electrical and Electronics Engineers Xplore (IEEE) , Vol.8, Issue 26 December 2020| ISSN: 38090-38109
- [7] Dr.Aziz Makandar, " Image Caption Generator Using CNN-LSTM", International journal of Advances in Engineering and Management , vol 4, issued on 10 October 2022.
- [8] Chetan Amritkar, Vaishali jabade, " Image Caption Generation Using Deep Learning Technique", Institute of Electrical and Electronics Engineer (IEEE) , Vol no 7, issued on November 2019 |ISSN : 172311 172338
- [9] Grishma Sharma, "Visual Caption Generator using Deep Learning" 2nd International Conference on Advances in Science and Technology , issued on 29 December 2019.
- [10] Kunal Vispute, "Image Caption Generator Using Deep Learning Approach", International Journal of Advanced Research in Science , Communication and Technology, vol 3, issued on 6 May 2023.
- [11] Reshmi Sasibhooshan, Suresh Kumaraswamy "Visual Caption Generation Using Visual Attention Prediction and Contextual Spatial Relation Extraction", issued on 18 February 2023.
- [12] Mr. Dharendra Parate, Mrs. Minu Choudhary, "Image Caption Generator using deep learning with Flickr Dataset", issued on 2022 IJRTI | Volume 7, Issue 8 |
- [13] Mr. S. K. S. Ibrahim, Mr. S. S. Kumar, "Apply Deep Learning-based CNN and LSTM for Visual Image Caption Generator", International Conference on Applied Artificial Intelligence and Computing (ICAAIC), IEEE, Volume: 2023, Published: June 2023.
- [14] Chaithra V, Charitra Rao, Deeksha K "IMAGE CAPTION GENERATOR USING DEEP LEARNING", International Journal of Engineering Applied Sciences and Technology, issued on June 2022.
- [15] Palak Kabra, Mihir Gharat, "Image caption generator using deep learning", International Journal of Advanced Research in Science, Engineering and Technology , Volume: 2022, published: October 2022 .
- [16] Antonio M. Rinaldi, Cristiano Russo, "Automatic image captioning combining natural language processing and neural networks", Information Sciences, Volume 648, issued on October 2023.
- [17] Farida Attar1, Farzana Khan, Affan Ansari, Mujawar Saklen, Abubakr Shaikh, Danish Khan , "Image Caption Generator using Deep Learning ", International Journal of Advanced Research in Science, Communication and Technology (IJARSCT), Volume 4, Issue 7, April 2024 .
- [18] Aryan Chauhan, , Miss Diksha Arya, " Image Caption Generator Using Deep Learning", Journal of Emerging Technologies and Innovative Research (JETIR), Published on April 2023, Volume 10, Issue 4
- [19] T. Sandhya, Dr. Kondapalli Venkatesh "GENERATING IMAGE CAPTIONS BASED ON DEEP NEURAL NETWORKS ", International Journal of Novel Research and Development, Volume 8, Issue 9 September 2023 | ISSN: 2456-4184 |
- [20] Mr. M. Bhalekar, Mr. M. Bedekar, "D-CNN: A New Model for Generating Image Captions with Text Extraction Using Deep Learning for Visually Challenged Individuals", Engineering, Technology & Applied Science Research (ETASR), Volume 12, Issue 2, April 2022, Pages 8366–8373, ISSN: 2241 4487, eISSN: 1792-8036