IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

AI BASED IMAGE CAPTION GENERATOR USING CNN

Author(s): Shabira S*[1], Venkatesan C*[2], Manjunath S*[3], Prathap M*[4]

Designation: Assistant Professor*[1], Students's*[234]

Department Of Computer Science And Engineering

Adhiyamaan College Of Engineering, Hosur, Tamil Nadu, India.

ABSTRACT

Image captioning is a crucial task in artificial intelligence that bridges the gap between computer vision and natural language processing. This project, AI-Based Image Caption Generator, aims to generate accurate and meaningful captions for images using deep learning techniques. The system integrates a Convolutional Neural Network (CNN) for feature extraction and a Recurrent Neural Network (RNN) with Long Short-Term Memory (LSTM) for sequential caption generation. To improve caption relevance, an attention mechanism is employed, allowing the model to focus on key image regions while generating descriptions. The model is trained on benchmark datasets like MS COCO and Flickr8k/30k, ensuring robustness in diverse image scenarios. Performance evaluation is conducted using standard metrics such as BLEU, METEOR, and CIDEr to assess the quality of generated captions. The proposed system has significant applications in image-based search, aiding visually impaired individuals, and automated content generation. This project aims to enhance the accuracy and contextual relevance of image descriptions, contributing to advancements in AI-driven image understanding.

Key Words: Artificial intelligence, Convolutional Neural Network, Image Caption Generator, Recurrent Neural Network, LSTM, CIDEr, MS COCO.

I.INTRODUCTION

AI-based image captioning is a powerful deep learning application that generates textual descriptions for images by combining computer vision and natural language processing. This system bridges the gap between visual perception and linguistic understanding, enabling machines to interpret and describe images in a human-like manner. It has significant implications for accessibility, content organization, and automated metadata generation. This system is built using Convolutional Neural Networks (CNNs) for extracting rich visual features from images and Recurrent Neural Networks (RNNs), specifically Long Short-Term Memory (LSTM) units, for generating coherent and contextually relevant captions. The CNN component processes the image to identify key objects, textures, and spatial relationships, transforming them into a structured feature representation. These extracted features are then fed into the LSTM-based RNN, which sequentially generates descriptive text by predicting each word based on the learned context. To improve accuracy and fluency, advanced techniques such as attention mechanisms and transformer-based models can be integrated. Attention mechanisms enhance the model's ability to focus on the most relevant parts of an image when generating captions, while transformer architectures such as Vision Transformers (ViTs) and Large Language Models (LLMs) provide greater contextual awareness and linguistic diversity. Training an image captioning model requires large, annotated datasets such as MS COCO (Microsoft Common Objects in Context), Flickr30K, and Conceptual Captions, which contain thousands of images paired with human-generated descriptions. The model learns from these datasets using supervised learning techniques, optimizing performance through sequence-to-sequence learning, cross-entropy loss, and reinforcement learning strategies such as Self-Critical Sequence Training (SCST) to generate more natural and accurate captions. This technology has diverse applications, including: Assistive Technology for the Visually Impaired – Converts images into spoken descriptions, enhancing accessibility. Automated Image Tagging and Organization – Helps in efficient image classification and retrieval. Enhanced Image Search and Indexing – Improves search engine accuracy by adding contextually relevant metadata. Social Media and Content Creation – Automates caption generation for posts, reducing manual effort. Surveillance and Security – Analyzes and describes surveillance footage for better monitoring and anomaly detection. By integrating cutting-edge deep learning techniques, this project aims to develop a highly efficient and accurate AI-driven image captioning system that can produce human-like, context-aware descriptions, ultimately enhancing image comprehension, accessibility, and automation across various domains.

1.1 CONVOLUTIONAL NEURAL NETWORK

A specific kind of deep learning model called a Convolutional Neural Network (CNN) is used to process and analyze visual input, especially photos and videos. It takes its cues from how the visual cortex of an animal is structured, which enables it to automatically learn from and extract information from unprocessed input data. For several computer vision applications, including object identification, picture segmentation, and image classification, CNNs have emerged as the cutting-edge method. Convolutional layers, pooling layers, and fully linked layers are the main parts of a CNN. In order to extract features and

produce feature maps, convolutional layers apply filters to the input data. By reducing the spatial dimensions of the data, pooling layers efficiently lower the computational load and manage overfitting. The final predictions or classifications based on the learnt features are handled by fully connected layers. CNNs can identify objects and features in a picture independent of their position because of their capacity to learn local patterns and spatial hierarchies. Because of this, they are very good at jobs requiring the comprehension of intricate visual patterns. CNNs' hierarchical architecture makes it possible for them to reliably and accurately recognize objects and shapes from low-level features like edges and textures to high-level features like shapes.

1.2 MS COCO

MS COCO (Microsoft Common Objects in Context) is one of the most widely used datasets for image captioning, object detection, and segmentation. It contains richly annotated images with multiple captions per image, making it ideal for training deep learning models in computer vision and natural language processing (NLP).

1.3 CIDEr

CIDEr (Consensus-based Image Description Evaluation) is a metric used to evaluate the quality of generated image captions by comparing them with multiple human-annotated reference captions. It measures how well the generated caption matches human descriptions using TF-IDF weighting and n-gram similarity.

CIDEr Formula

The CIDEr score is computed as:

$$CIDEr = \frac{1}{N}\sum_{n=1}^{N} w_n \cdot CIDEr_n$$

II. EXISTING SYSTEM

Existing image captioning systems have evolved from simple encoder-decoder architectures to advanced attention-based and transformer-driven models. Early systems relied on CNNs for feature extraction and RNNs or LSTMs for text generation, as seen in models like Show and Tell. Attention-based approaches, such as Show, Attend and Tell, improved caption relevance by dynamically focusing on important image regions. Transformer-based models like Image Transformer and multimodal learning approaches like Oscar further enhanced captioning by capturing complex dependencies and integrating textual and visual data. More recent models, including BLIP and Flamingo, leverage large-scale pretraining and multimodal fusion for high-quality captions. Despite these advancements, existing systems struggle with uncommon objects, bias issues, and a lack of interpretability, necessitating further improvements in attention mechanisms, bias reduction, and interactive learning. Recent advancements also explore contrastive learning techniques to improve caption diversity and contextual accuracy. Interactive learning strategies are being

investigated to refine captions based on real-time user feedback. Additionally, efforts are being made to enhance explainability using attention maps and visualization techniques to provide better insights into model decision-making.

III. PROPSED SYSYEM

The proposed image captioning system utilizes the MSCOCO dataset for training and CIDEr as a key evaluation metric to enhance caption quality. It integrates advanced attention mechanisms, such as self-attention and multi-head attention, to focus on significant image regions dynamically**.** Multimodal fusion techniques, including cross-modal attention and early-late fusion, improve the interaction between visual and textual data. Contrastive learning **i**s used to pretrain on large datasets, enhancing the model's ability to handle rare objects and complex scenes. Context-aware captioning ensures captions are descriptive and meaningful by analyzing contextual cues within images. To address biases, techniques like data augmentation, domain adaptation, and bias correction are implemented. The system also incorporates interactive learning**,** allowing user feedback to refine and improve captions over time. Evaluation metrics, including CIDEr, BLEU, and METEOR, are combined with human-centric measures like coherence and relevance. Attention maps and visualization techniques improve interpretability by highlighting image regions influencing caption generation. This approach ensures more accurate, diverse, and contextually aware captions**,** making the system more reliable and user-driven.

System Overview

The proposed image captioning system is designed to generate high-quality, context-aware captions by leveraging MSCOCO as the primary dataset and CIDEr for evaluation. It integrates advanced attention mechanisms like self-attention and multi-head attention to dynamically focus on significant image regions, improving caption relevance**.** Multimodal fusion strategies**,** including cross-modal attention and early-late fusion, ensure seamless integration of visual and textual features. Contrastive learning enhances caption accuracy by pretraining on large datasets, making the model more robust in recognizing rare objects. The system also incorporates context-aware captioning, which uses visual and semantic cues to generate meaningful and descriptive captions. Additionally**,** interactive learning allows users to provide feedback, refining the model's performance over time. Bias reduction techniques, such as data augmentation and domain adaptation, ensure fairness in caption generation. The system's outputs are evaluated using CIDEr, BLEU, and METEOR, along with human-centric measures like coherence and relevance. Attention visualization techniques further improve interpretability, providing insights into how the model generates captions.

Benefits

- **Improved Caption Accuracy** – Advanced attention mechanisms ensure precise and contextually relevant captions.
- **Better Visual-Text Integration** – Multimodal fusion strategies enhance the relationship between image features and textual descriptions.
- **Enhanced Robustness** – Contrastive learning and large-scale pretraining improve handling of rare objects and noisy data.
- **User-Driven Refinement** – Interactive learning allows users to provide feedback for more accurate and personalized captions.
- **Comprehensive Evaluation** – CIDEr, BLEU, METEOR, and human-centric metrics ensure a well-rounded assessment of caption quality.

IV.METHODOLOGY

The methodology of the proposed image captioning system begins with data preprocessing using the MSCOCO dataset, where images and corresponding captions are cleaned and tokenized**.** A CNN-based visual encoder extracts image features, which are then processed by a transformer-based decoder for text generation. Advanced attention mechanisms, including self-attention and multi-head attention, help focus on crucial image regions for better captioning. Multimodal fusion techniques, such as cross-modal attention and early-late fusion, enhance the integration of visual and textual information. Contrastive learning is applied during pretraining to improve robustness against rare objects and noisy data. The model undergoes fine-tuning with feedback-driven adjustments through interactive learning, allowing users to refine captions. Bias reduction techniques**,** such as domain adaptation and data augmentation, ensure fair and inclusive caption generation. The system is evaluated using CIDEr, BLEU, and METEOR, along with human-centric metrics like coherence and relevance. Attention visualization techniques improve interpretability by highlighting key image areas influencing caption decisions. This methodology ensures accurate, diverse, and context-aware captions, making the system more effective and user-adaptive.

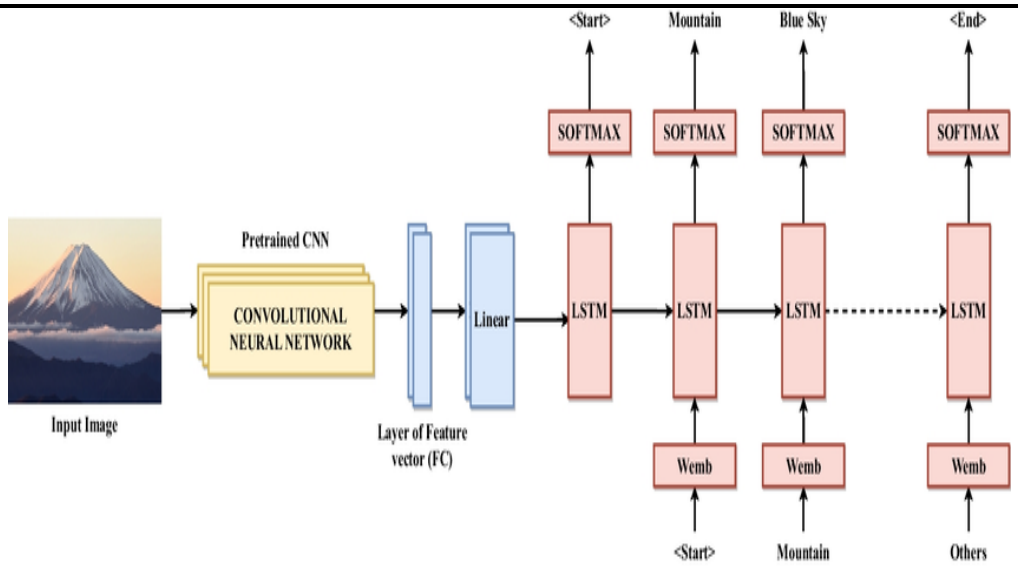

Figure 1: System Architecture

V. PROPOSED MODULE AND ALGORITHM

5.1 LSTM (LONG AND SHORT TERM MEMORY)

Long Short-Term Memory (LSTM) networks, a specialized type of RNN, address the vanishing gradient problem and effectively capture long-term dependencies in sequential data. They are widely used in NLP, time series prediction, and image captioning. In image captioning, a pre-trained CNN like VGG or ResNet extracts high-level image features, which are then processed by an LSTM to generate meaningful captions. During training, the model learns from image-caption pairs, using CNN-derived features and previously generated words to predict the next word in the sequence. At inference, the CNN extracts features from a new image, and the LSTM sequentially generates captions until an end token is reached. Activation functions like ReLU and Softmax optimize learning for better accuracy. The combination of CNNs for feature extraction and LSTMs for sequence modeling enables more coherent and context-aware captions. This approach enhances caption quality by effectively linking visual and textual data.

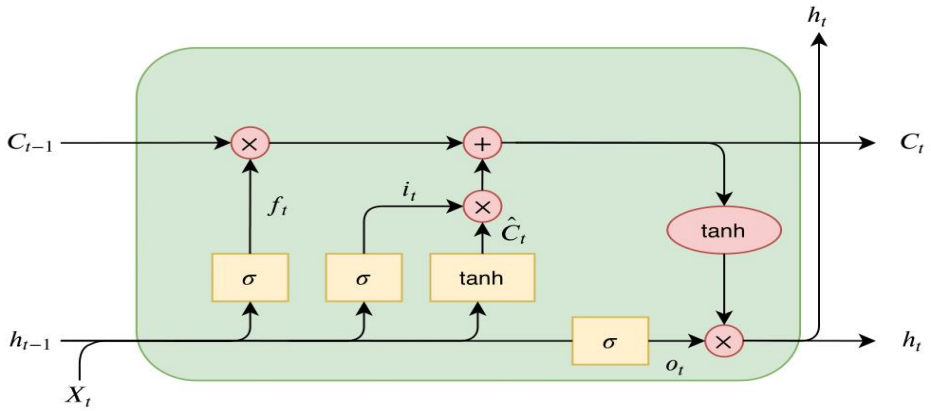

Figure 2: Working Of LSTM

5.2 ARCHITECTURE/WORKFLOW

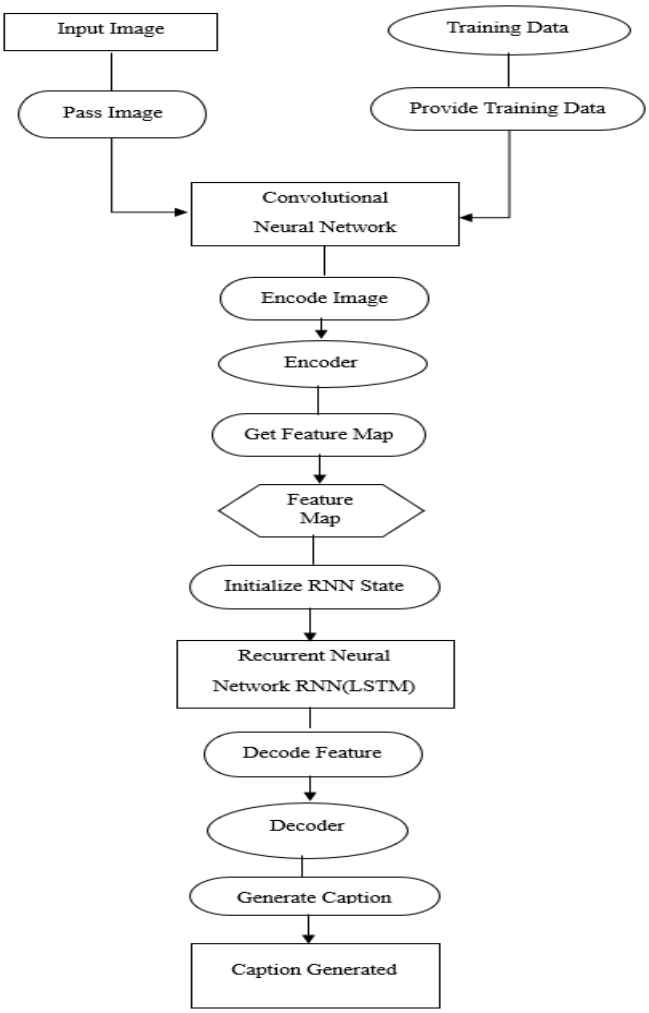

Figure 3: Workflow Diagram

VI. RESULTS AND DISCUSSION

EXCEPTED OUTCOME

- **Accurate and Context-Aware Captions** – The system will generate precise, meaningful, and contextually relevant captions by integrating advanced attention mechanisms and multimodal fusion techniques.
- **Improved Handling of Rare Objects and Complex Scenes** – Contrastive learning and pretraining on large datasets will enhance the model's ability to recognize and describe uncommon objects with greater accuracy.
- **User-Driven Refinement and Bias Reduction** – Interactive learning will allow users to provide feedback, improving caption accuracy over time, while bias reduction techniques will ensure fair and inclusive descriptions.

- **Enhanced Interpretability and Evaluation** – Attention visualization techniques will provide insights into the caption generation process, while evaluation metrics like CIDEr, BLEU, and METEOR will ensure comprehensive performance assessment.

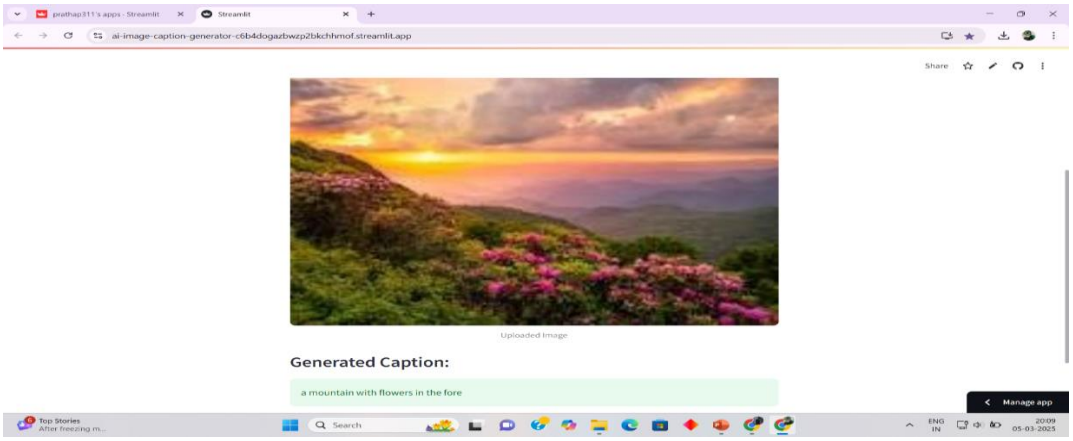

Figure 4: Generated Output

VII. CONCLUSION

The AI-Based Image Caption Generator effectively combines CNNs for feature extraction and LSTMs for sequential text generation to produce accurate and contextually relevant image descriptions. Advanced attention mechanisms and multimodal fusion techniques enhance the model's ability to focus on key image regions, improving caption quality. Contrastive learning and interactive feedback further refine caption accuracy while addressing biases. The system's evaluation using CIDEr, BLEU, and METEOR ensures a comprehensive performance assessment. With applications in accessibility, content automation, and image retrieval, this project contributes to AI-driven image understanding. Future enhancements may include transformer-based architectures and real-time user interaction for even more refined captioning.

VIII. REFERENCES

[1] O. Vinyals et al., "Show and Tell: A Neural Image Caption Generator," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2015.

[2] A. Vaswani et al., "Attention is All You Need," *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[3] P. Anderson et al., "Bottom-Up and Top-Down Attention for Image Captioning and VQA," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[4] K. He et al., "Deep Residual Learning for Image Recognition," *IEEE CVPR*, 2016.

[5] J. Devlin et al., "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *Proceedings of NAACL-HLT*, 2019.

[6] T. Ghandi, H. Pourreza, and H. Mahyar, "Deep Learning Approaches on Image Captioning: A Review," arXiv preprint arXiv:2201.12944, 2022.

[7] P. Bhatnagar, S. Mrunaal, and S. Kamnure, "Enhancing Image Captioning with Neural Models," arXiv preprint arXiv:2312.00435, 2023.

[8] Y. Wang et al., "Deep image captioning: A review of methods, trends and future challenges," Neurocomputing, vol. 546, pp. 126287, 2023.

[9] M. S. R. S. K. S. V. S. A. S. S. "Image caption generation using Visual Attention Prediction and Contextual Spatial Relation Extraction," Journal of Big Data, vol. 10, no. 1, pp. 1-24, 2023.