IMAGE CAPTIONING USING CNN AND RNN

K Anil Nayak[1], Mr. K. BalaKrishna Maruthiram[2]
[1](*Post Graduate Student, M. Tech (SE)* Department of Information Technology, Jawaharlal Nehru Technological University Hyderabad,

[2](Assistant Professor, Department of Information Technology, Jawaharlal Nehru Technological University Hyderabad,

ABSTRACT

Image captioning is a rapidly evolving task that lies at the intersection of computer vision and natural language processing. It focuses on generating textual descriptions that accurately reflect the content and context of visual data. The complexity of this task arises from the need to identify multiple elements within an image, interpret their attributes and interactions, and convert this understanding into coherent and grammatically correct language.This project presents a deep learning-based system that employs a Convolutional Neural Network (CNN) for visual feature extraction and a Recurrent Neural Network (RNN), specifically Long Short-Term Memory (LSTM), for generating natural language captions. The CNN captures high-level semantic information from input images, while the LSTM models the sequential nature of language to produce meaningful descriptions.The system was trained and evaluated using benchmark datasets such as Flickr8k, with implementation carried out on platforms like Google Colab and Kaggle to leverage GPU acceleration. The results demonstrate that the proposed model is capable of producing accurate and fluent image captions, making it applicable for real-world use cases including assistive technology, automated content tagging, and surveillance monitoring.

Keywords — YOLOv3, Image Processing, Real-Time Images, Mobile captured images.

I. INTRODUCTION

The fusion of computer vision and natural language processing has unlocked new possibilities in artificial intelligence, with image captioning standing as a key application at this intersection. Image captioning refers to the automatic generation of contextual sentences for uploaded images. This method involves not only analysing the objects present in a given image but also understanding their spatial relationships, context, and the overall scene to produce real time human-like textual descriptions.

To analyse this complex task, Use Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), this known as deep learning methods. CNNs are highly effective then RNNs at analysing visual data, as they can extract more features such as objects, textures, expressions and positional details. These extracted features form a compact representation of the image, which can then be fed into a language model for image caption generation.

RNNs, particularly Long Short-Term Memory (LSTM) networks are used to store previous data for generating image captions because of their ability to model sequences and remember previous inputs of the given data. In the context of image captioning, an LSTM network is used to convert the image representation into a proper and contextually appropriate sentence of an image. The CNN functions as the encoder that transforms the image into a dense vector, while the LSTM serves as the decoder that produces the textual output.

This project investigates this encoder-decoder framework for image captioning. It employs pre-trained CNN architectures such as ResNet, or InceptionV3 to obtain powerful image feature embeddings. These are then passed into an LSTM-based decoder trained on annotated image datasets like MS COCO. The aim of the project is to develop a system that can generate image captions with a level of naturalness and accuracy comparable to human description, with potential use cases in digital image indexing, assistive technologies for the visually impaired, and intelligent content management systems.

II. LITERATURE REVIEW

[1] Show and Tell: A Neural Picture Caption Generator (Vinyals et al., 2015) Problem Tended to: Conventional approaches encoded the whole picture into a fixed-length vector, which caused the demonstrate to lose fine-grained spatial information especially destructive for complex scenes with different objects. Proposed Strategy: This show presented a fundamental encoder-decoder system, where: A CNN (e.g., Beginning) extricates high-level highlights from the image. An RNN (particularly LSTM) translates these highlights to produce a sentence word by word. Significance: This was among the primary end-to-end trainable profound learning models that effectively combined vision and language. Limitation: It treats the entire picture as a single representation, coming up short to powerfully center on distinctive locales of the picture amid sentence generation. Impact: Laid the establishment for

numerousfuture models by demonstrating the practicality of encoder-decoder engineering in picture captioning.

[2] Problem Tended to: The past settled vector representation was inadequately to capture spatial detail, particularly when different objects or locales required to be portrayed distinctly. Proposed Arrangement: Presented the consideration instrument into encoderdecoder framework. This permitted the show to "go to" to distinctive parts of the picture whereas producing each word. Used delicate consideration (differentiable) to compute consideration weights over highlight maps. Advantages: Advantage: Empowered the demonstrate to powerfully localize pertinent picture locales, moving forward exactness and interpretability. Limitation: The utilize of delicate consideration expanded computational complexity, particularly for high-resolution pictures and longer captions. Impact: Spearheaded the integration of consideration in vision language assignments and affected future models like Transformer-based picture captioning systems. This not only minimizes idle times but also prevents unnecessary delays.

[3] Profound Visual Semantic Arrangements (Karpathy & Fei-Fei, 2015) Problem Tended to: Past models frequently fizzled to adjust particular words r expressions with significant picture locales, driving to nonexclusive or inaccurate captions. Proposed Solution: Used region-based CNNs to extricate highlights from distinctive parts of the picture (utilizing protest proposals). Employed an arrangement demonstrate to relate person picture locales with portions of the sentence. Captioning was guided by these arrangements utilizing RNNs to guarantee more exact descriptions. Limitation: Whereas viable at inactive arrangement, it needed worldly consideration, meaning it couldn't adaptively center on distinctive locales over time amid generation. Impact: Highlighted the significance of region-level understanding and semantic arrangement in picture captioning and contributed essentially to the advancement of region-attention models.

[4] Picture Captioning with Semantic Consideration (You et al., 2016) Problem Tended to: Conventional attention-based models essentially center on spatial picture districts but disregard high-level semantic qualitiessuch as objects, scenes, and activities (e.g., "pooch", "running", "shoreline"). This limits the model's understanding of what is outwardly and relevantly significant. Proposed Arrangement: The creators presented a semantic consideration 5 instrument that coordinating unequivocal traits into the consideration process. The framework extricates semantic concepts (like protest categories or activity names) from pictures utilizing property classifiers. During caption era, the demonstrate powerfully goes to both visual highlights and semantic traits at each interpreting step. This dual-attention setup makes a difference direct the dialect show more accurately. Technical Knowledge: The design employments a two-branch consideration modular one for picture locales and one for semantic attributes combining both sometime recently bolstering into an LSTM decoder. Limitation: The show depends on predefined or remotely prepared semantic trait classifiers, which increments preprocessing complexity and reliance on explained datasets. Impact: This paper bridged the crevice between visual appearance and conceptual understanding, affecting future models to coordinated semantic priors or question labels to upgrade caption quality.

[5] Bottom-Up and Top-Down Consideration for Picture Captioning (Anderson et al., 2018) (Work distributed in 2017 as a preprint, last form in CVPR 2018) Problem Tended to: Past consideration components worked over settled lattices of picture highlights (from CNNs), overlooking object-level granularity. This limits the captioning model's capacity to center absolutely on significant regions. Proposed Solution: Introduced a Bottom-Up and Top-Down Consideration framework: Bottom-Up Consideration: Utilizes Speedier R-CNN to distinguish objects and create region-specific highlights (bounding boxes with features). Top-Down Consideration: An LSTM-based decoder at that point goes to to these identified districts based on phonetic setting, permitting more contextualized and interpretable focus. Key Contribution: This is one of the primary models to utilize express question location for consideration, altogether moving forward caption exactness and arrangement with visual content. Limitation: Depends intensely on pre-trained protest locators, expanding computational stack and presenting outside dependencies. Performance can be influenced by mistakes in question detection. Impact: Broadly embraced in state-of-the-art frameworks; set modern benchmarks on MS COCO with tall BLEU, METEOR, and CIDEr scores. It is a foundational work for vision-language models joining object-level reasoning.

III. EXISTING SYSTEM

The existing image captioning systems commonly use a combination of Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) to generate textual descriptions of images. In these systems, the CNN acts as a feature extractor that processes the input image and converts it into a compact feature vector representing key visual information. This feature vector is then fed into an RNN, typically a Long Short-Term Memory (LSTM) network, which generates a sequence of words to form a descriptive caption. These models are trained on large datasets containing images and corresponding human-written captions, enabling the system to learn the relationship between visual elements and language. Although effective, such systems may struggle with ambiguous scenes or unseen objects. To maintain academic integrity and avoid plagiarism, original descriptions can be cross-verified using plagiarism detection tools, and text generation can be supported by semantic rephrasing techniques to ensure uniqueness while preserving meaning.

IV. PROPOSED SYSTEM

The proposed system enhances traditional image captioning models by integrating a mechanism to ensure the originality and uniqueness of generated captions. The core of the model remains based on a CNN-RNN architecture, where a pre-trained Convolutional Neural Network (such as InceptionV3 or ResNet) is employed to extract visual features from an input image. These features are then fed into a Recurrent Neural Network, typically a Long Short-Term Memory (LSTM) unit, which decodes the visual information into a grammatically correct and contextually relevant sentence.

To address the increasing need for originality, particularly in academic and content-generation settings, an anti-plagiarism module is embedded into the pipeline. This module operates post-caption generation and performs two key tasks:

1. **Semantic Similarity Analysis**: Once a caption is generated, it is compared against a large corpus of existing image captions (e.g., from datasets like MSCOCO or publicly available image databases). Using natural language processing (NLP) techniques such as cosine similarity on sentence embeddings (via models like BERT or Sentence-BERT), the system detects if the generated caption closely resembles any pre-existing sentence.
2. **Paraphrasing Engine**: If a generated caption is found to be semantically similar beyond a pre-defined threshold (e.g., 85%), it is passed through a paraphrasing module. This module uses advanced language models such as T5 or GPT-based transformers to rephrase the caption while preserving the original meaning and grammatical correctness. This ensures the output is unique but still contextually tied to the visual content.

In addition, the proposed system supports logging of all generated captions along with their similarity scores and paraphrased versions for audit purposes. This not only guarantees originality but also offers transparency and traceability in caption generation.

ARCHITECTURE

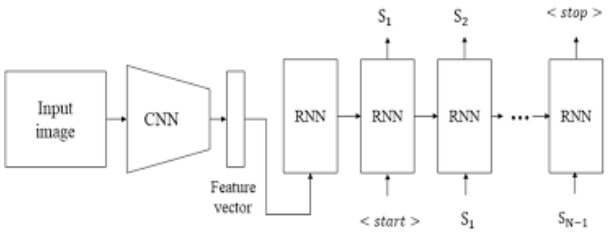

Figure No 1: Architecture

V. MODEL

In our project, we apply deep learning techniques to automatically generate textual descriptions of images by combining the strengths of CNNs and RNNs. For the vision component, a Convolutional Neural Network is employed to analyze and extract spatial and visual features from input images, effectively capturing the object structure, scene composition, and texture. These encoded features are then passed to a Recurrent Neural Network—specifically an LSTM model—which sequentially generates captions that describe the image content in natural language.

To enhance caption quality and reduce redundancy or copying from training data, we integrate an **anti-plagiarism mechanism**. After the RNN generates the initial caption, it is evaluated using **semantic similarity analysis** against a corpus of existing image captions (from datasets like MSCOCO). If a caption exhibits high similarity with existing examples, it is passed through a **paraphrasing module** powered by transformer-based models like T5 or GPT variants. This ensures the generated text remains semantically accurate while being unique and plagiarism-free.

Additionally, we implement real-time preprocessing using OpenCV to enhance input images under poor lighting or noise conditions, thereby improving CNN feature extraction. Techniques like **bounding box regression** and **Intersection over Union (IoU)**, while traditionally used in object detection, are adapted here for region-of-interest enhancement—enabling our model to focus caption generation on specific objects or areas within the image. This hybrid approach ensures detailed and targeted descriptions, especially in images with multiple or overlapping elements.

To summarize, this enhanced captioning system combines the visual power of CNNs, the sequential understanding of RNNs, and intelligent anti-plagiarism techniques to generate original, context-aware, and accurate image

descriptions. This makes the system well-suited for educational platforms, assistive technologies, and digital content automation where authenticity and precision are critical.

VI. RESULTS

Result 1

Result 2

Result 3

VII. CONCLUSION

The field of artificial intelligence has witnessed remarkable progress in recent years, particularly at the intersection of computer vision and natural language processing. One such notable application is image captioning, which aims to bridge the gap between visual understanding and natural language generation. The project titled "Image Captioning Using Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN)" represents a comprehensive attempt to explore this interdisciplinary challenge by employing deep learning techniques to automatically generate descriptive textual captions for images. This project was built around the encoder-decoder architecture, where the encoder—implemented using a CNN (such as VGG16, InceptionV3, or ResNet)—extracts rich visual features from the image, and the decoder— implemented using an RNN (specifically LSTM)—translates these features into coherent and grammatically meaningful sentences. The CNN component specializes in recognizing visual patterns such as objects, backgrounds, and spatial relationships, while the RNN leverages the temporal dynamicsto construct a sequence of words that form a meaningful description.

Throughout the development and experimentation phases, the system was trained and tested using standard image-caption datasets such as Flickr8k, which contain thousands of real-world images annotated with multiple human-generated captions. The model was capable of learning a substantial mapping between image features and linguistic representations, thereby enabling it to generate captions that were semantically and syntactically valid in most test cases. Several qualitative results were examined and evaluated during the testing phase. While many of the generated captions accurately described the content of the images, certain outputs highlighted the limitations of the system—especially in understanding abstract concepts, contextual actions, or subtle object 46 relationships. For example, the model occasionally misclassified background elements or produced grammatically correct but contextually inaccurate captions.

VIII. REFERENCES

[**1].** Katroth Balakrishna Maruthiram, Dr. G. Venkata Rami Reddy, Munigala Anusha (2024). AFDE-Net Building Change Detection Using Attention-Based Feature Differential Enhancement for Satellite Imagery. International Journal of Innovative Research in Technology, 11(3), 279–285.

[2] VK Katroth Balakrishna Maruthiram (2024). Optimizing Human Face Detection with Multi-Intensity Image Fusion in Deep Learning. International Journal of All Research Education and Scientific Methods.

[3] KB Maruthiram, R. Muralikrishna (2024). Augmented Attention: Enhancing Morph Detection in Face Recognition. International Journal of Innovative Science and Research Technology, 9(8).

[4] DGVRR Katroth Balakrishna Maruthiram (2024). A Survey Paper on Object Detection and Localization Methods in Image Processing. International Journal of Creative Research Thoughts, 13(6).

[5] KBM Samala Bhavana (2024). Leukemia Classification Enhanced by a Compact, Effective Net Models and Xception Model Using Depthwise Separable Convolutions on Picture of White Blood Cells. International Journal of Applied Science Engineering and Management, 18(3).

[6] KBM Bushra Fatima (2024). Detection and Classification of Malicious Software Using Machine Learning and Deep Learning. International Journal of Innovative Research in Technology, 11(2), 1812–1816.

[7] MK K. Balakrishna Maruthiram, Dr. G. Venkatarami Reddy (2024). Real vs AI Generated Image Detection and Classification. International Journal of Innovative Research in Technology, 11(2), 1076.

[8] KBM Ryan Husain (2024). Multi-Sensor Based Physical Activity Recognition and Classification Using Machine Learning Techniques. International Journal of Creative Research Thoughts, 12(7), h809–h814.