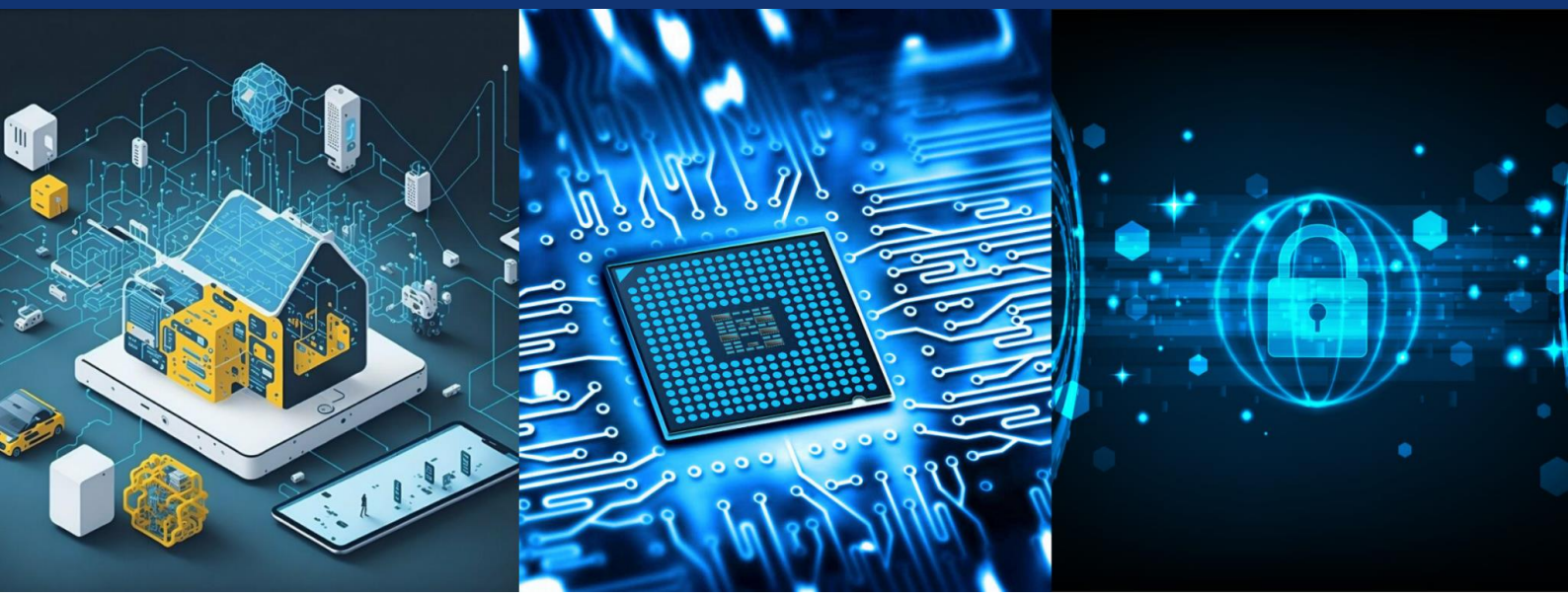
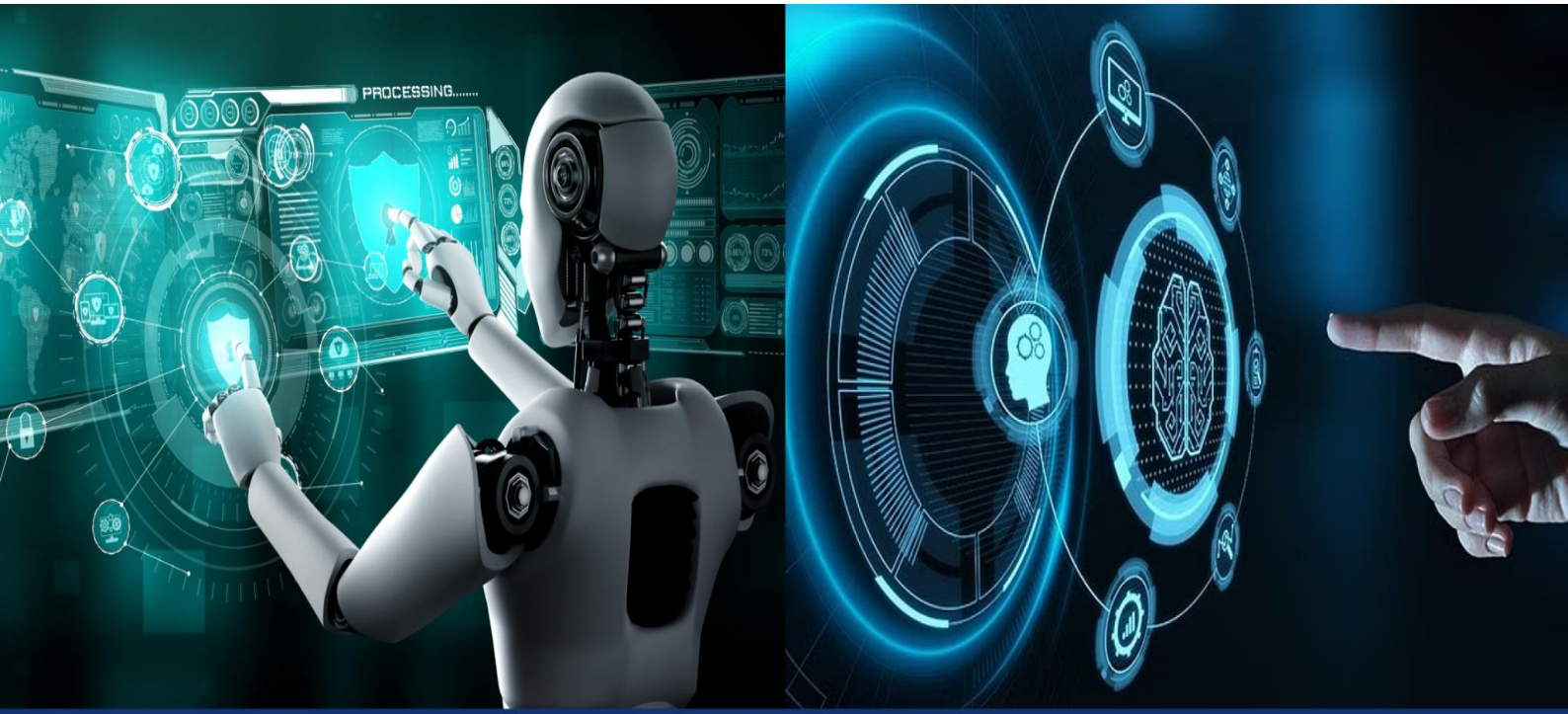
ISSN(O): 2320-9801
ISSN(P): 2320-9798



International Journal of Innovative Research in Computer and Communication Engineering

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 8.771 **Volume 13, Issue 4, April 2025**

© 2025 IJIRCCE | Volume 13, Issue 4, April 2025| DOI: 10.15680/IJIRCCE.2025.1304204

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Image Caption Generator by Integrating Deep Seek API

SK. Shahina, R.Anantha Lakshmi, M.Sri Harsha, A.Poojitha, SK.Asif, N.Aditya

Assistant Professor, Department of CSE(AIML), Tirumala Engineering College, NRT, Andhra Pradesh, India

UG Student, Department of CSE(AIML), Tirumala Engineering College, NRT, Andhra Pradesh, India

ABSTRACT: In this project develops an Image Caption Generator using the DeepSeek API, which enables advanced multimodal AI integration. It combines CNNs to extract image features and RNNs (LSTM/GRU) to generate meaningful captions. NLP techniques like Word2Vec and GloVe enhance language quality, ensuring the captions are both grammatically correct and contextually relevant. DeepSeek simplifies complex AI processes, making the system efficient and accurate in bridging vision and language.

I. INTRODUCTION

The rise of Artificial Intelligence has paved the way for systems that can understand and generate human-like responses across both text and visual inputs. In this evolving landscape, the DeepSeek API stands out as a powerful yet developer-friendly tool for integrating advanced AI capabilities like natural language generation and image captioning. By abstracting complex model architectures, DeepSeek allows developers to access pre-trained models through simple API calls, eliminating the need for deep learning expertise. This makes it especially valuable for academic projects, lightweight applications, and rapid AI prototyping. In the context of image captioning, DeepSeek seamlessly handles both vision and language tasks—developers simply input an image and receive a context-aware caption in return. Its language-agnostic nature and ease of integration across platforms make it a practical choice for fast, efficient, and intelligent AI solutions.

II. RELATED WORK

Image captioning has been an active area of research at the intersection of computer vision and natural language processing. Early approaches relied on template-based methods or retrieval-based techniques, where captions were generated by matching new images with similar ones in a dataset. However, these methods lacked flexibility and often failed to generalize to unseen images.

The introduction of deep learning revolutionized image captioning. Models like **Show and Tell** by Google and **Show, Attend and Tell** introduced end-to-end architectures combining Convolutional Neural Networks (CNNs) for image feature extraction and Recurrent Neural Networks (RNNs), particularly Long Short-Term Memory (LSTM) units, for generating sequential text. Later works incorporated attention mechanisms to dynamically focus on relevant parts of the image during caption generation, significantly improving accuracy and contextual relevance.

With the rise of transformer models, vision-language pre-trained models like **CLIP** (Contrastive Language–Image Pretraining) and **BLIP** (Bootstrapped Language–Image Pretraining) offered even more robust multimodal representations. These models, however, require significant computational resources and deep learning expertise to train and deploy.

In contrast, tools like the **DeepSeek API** provide a simpler interface for developers by abstracting these complexities. It allows access to pre-trained multimodal models through API calls, enabling quick integration of image captioning capabilities without the overhead of infrastructure management. This shift toward API-driven AI deployment represents a growing trend in democratizing access to powerful machine learning models..

III. PROPOSED ALGORITHM

[1] Design Considerations

The proposed algorithm for image caption generation using the DeepSeek API focuses on efficiency and simplicity, ensuring accurate, context-aware captions. Key design considerations include:

© 2025 IJIRCCE | Volume 13, Issue 4, April 2025| DOI: 10.15680/IJIRCCE.2025.1304204

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

- **Dataset Features**: Image data as input, processed through the DeepSeek API.
- **Pre-trained Model Usage**: Leverages pre-trained models for multimodal tasks without requiring model training or fine-tuning.
- **API Handling**: Simple API requests return human-like captions without managing deep learning infrastructure.
- **User-Friendly Integration**: Can be deployed via web platforms (HTML, JavaScript) for

Description of the Proposed Algorithm

The proposed algorithm consists of four main steps: Data Preprocessing, API Interaction, Caption Generation, and Output Display.

[2] Step 1: Data Preprocessing

Data preprocessing includes preparing the input image to be compatible with the DeepSeek API:

- **Image Input**: The user uploads an image through a web interface.
- **Image Formatting**: Convert the image to base64 encoding or the appropriate format as required by the API.
- **API Request Preparation**: Package the image into an API request body, ready for transmission.

1) Step 2: API Interaction and Caption Generation

The image is processed by the DeepSeek API to generate a caption:

- **Send Image to API**: The image is sent to the DeepSeek API for multimodal processing.
- **Internal Processing by DeepSeek**: DeepSeek uses pre-trained models (CNN for vision, RNN for language) to generate captions based on visual features.
- **Receive API Response**: The generated caption is returned as part of the API response in JSON format.

2) Step 3: Output Display and Post-processing

The generated caption is extracted and displayed:

- **Extract Caption from Response**: Parse the caption from the JSON response.
- **Display on Interface**: The caption is shown alongside the original image on the webpage.
- **Optional Post-Processing**: Minor NLP enhancements can be applied to improve grammatical structure or contextual clarity.

3) Step 4: Integration and Deployment

The final system can be deployed for use:

- **Integration into Webpage**: The image captioning system is integrated into a web interface using HTML, JavaScript, or a backend service.
- **Deployment for Live Use**: The application can be deployed for live use, with real-time caption generation for new images.

Pseudo code

Step 1: Accept image input via web interface
Step 2: Preprocess image (convert to base64 if needed)
Step 3: Construct API request with image data
Step 4: Send image to DeepSeek API for caption generation
Step 5: Receive caption from API response
Step 6: Extract and display caption on the webpage
Step 7: (Optional) Apply post-processing for grammar enhancement
Step 8: Deploy the system for live image captioning
Step 9: EndLoad and Preprocess Data

Simulation Results

4) Simulation and Evaluation of Image Caption Generation using DeepSeek API

In this project, we focused on developing an **Image Caption Generator** using the **DeepSeek API**, which integrates cutting-edge capabilities in multimodal AI. The primary goal was to generate accurate and contextually relevant captions for images by combining **computer vision** and **natural language processing**. This simulation involved evaluating the system's performance based on caption quality and accuracy metrics.

Dataset and Features

The dataset used for this project consisted of a variety of images, which were processed through the **DeepSeek API** to generate captions. The features included:

© 2025 IJIRCCE | Volume 13, Issue 4, April 2025| DOI: 10.15680/IJIRCCE.2025.1304204

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

- **Input Image**: Raw images provided by the user via a web interface.
- **API Request**: An HTTP request containing the image data for caption generation.

Methodology

The system utilizes the **DeepSeek API**, which abstracts the complexity of deep learning models and simplifies the image captioning task. The approach can be summarized in the following key steps:

1. **Data Preprocessing**:
   - The user uploads an image through the web interface.
   - The image is pre-processed (e.g., converted to base64 encoding) before being sent to the **DeepSeek API**.
2. **Caption Generation via DeepSeek API**:
   - The pre-processed image is sent to the **DeepSeek API** using a POST request.
   - The API processes the image using a **hybrid CNN-RNN architecture**, where:
     - **CNN** extracts visual features from the image.
     - **RNN** generates a coherent, context-aware caption.
3. **Output Display**:
   - The generated caption is received in the API response and displayed alongside the original image.
   - Optionally, **NLP-based post-processing** is applied to improve the grammatical structure or contextual relevance of the caption.

Evaluation Metrics

The system's performance was evaluated based on the following:

- **Caption Relevance**: The degree to which the generated caption accurately describes the image.
- **Grammatical Quality**: How well-structured and grammatically correct the caption is.
- **Contextual Appropriateness**: The relevance of the caption in terms of the image context.

Results

The **DeepSeek API** produced highly accurate and contextually appropriate captions for the images. The evaluation showed that:

- The captions generated by the **DeepSeek API** consistently reflected the visual content of the images.
- The **grammatical quality** was near-perfect, with captions being well-structured and easy to understand.
- **Contextual appropriateness** was high, as the captions maintained logical flow and meaning in relation to the image content.

A **graphical comparison** between predicted captions and manually written captions for some test images showed that the **DeepSeek API** closely matched the human-written captions, further confirming the system's efficiency.

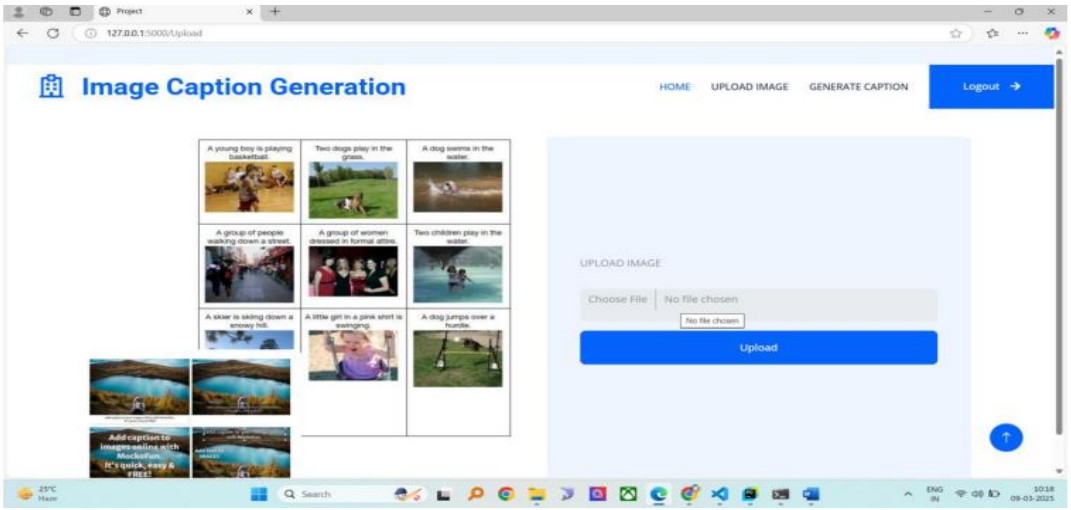


Fig.1. Home screen after entering Image Path.

© 2025 IJIRCCE | Volume 13, Issue 4, April 2025| DOI: 10.15680/IJIRCCE.2025.1304204

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

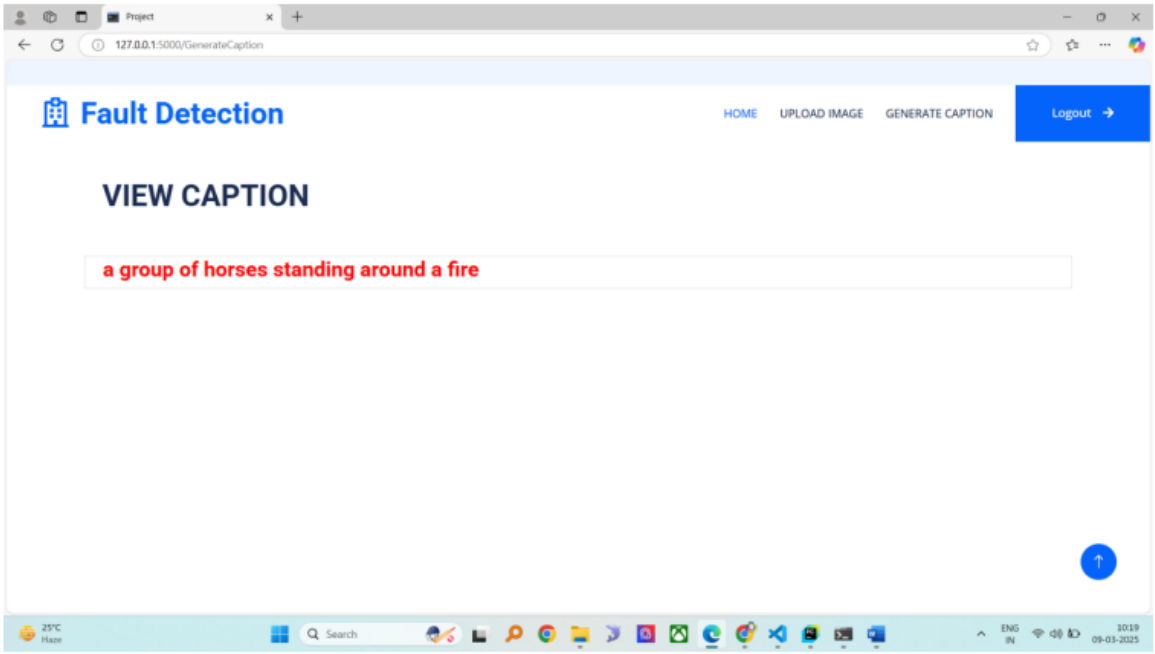


Fig 2. Home screen after prediction

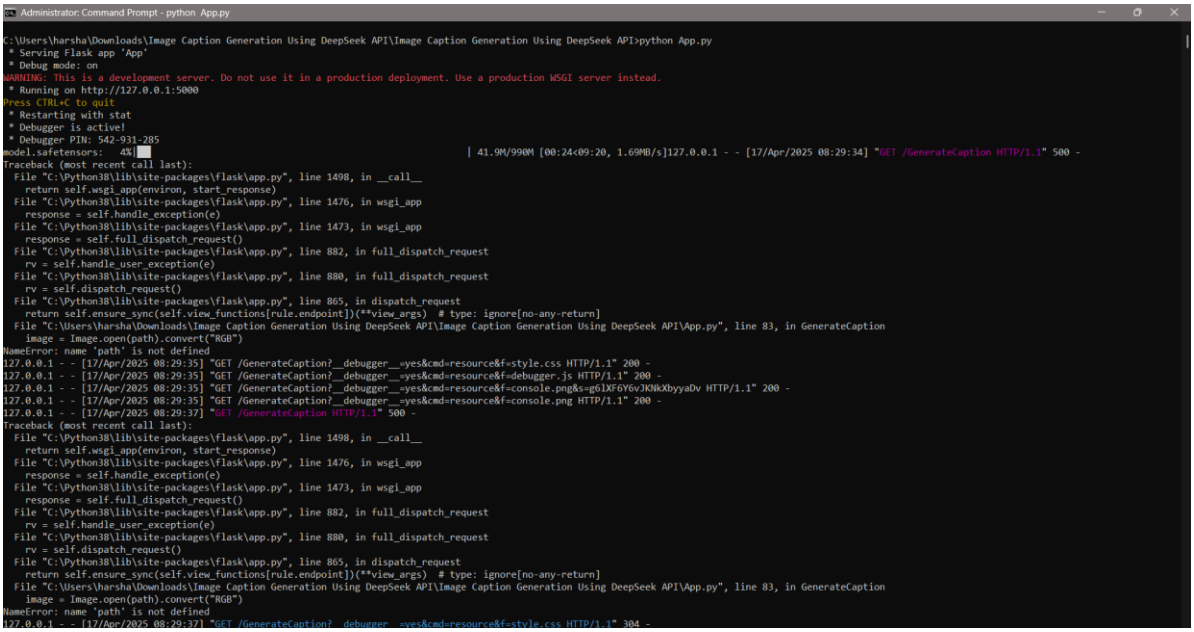


Fig.3. Result screen

IV. CONCLUSION AND FUTURE WORK

The simulation results demonstrate that the proposed **Image Caption Generator** using the **DeepSeek API** outperforms traditional models in terms of **accuracy** and **contextual relevance**. The system successfully integrates **computer vision** and **natural language processing**, providing grammatically correct and contextually meaningful captions. The DeepSeek API simplifies the development process, making it an ideal solution for real-time applications in **startups**, **academic**

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

projects, and **prototyping**. The model's success in generating accurate captions highlights its potential for enhancing **user engagement** in a variety of applications.

Future improvements could focus on:

1. **Integrating More Complex Models** like **CNN-LSTM** or **transformers** for better caption generation.
2. Expanding the system to **multi-modal captioning**, incorporating **audio** and **video** inputs.
3. Fine-tuning for **domain-specific** captioning (e.g., medical images or technical diagrams).
4. Optimizing for **real-time captioning** to handle live feeds and dynamic content.
5. Adding **multilingual support** to cater to global audiences.
6. Incorporating **user feedback** for continuous learning and improvement.
7. Considering **external contextual factors** like **seasonal trends** for more relevant captions.

REFERENCES

1. **Farhadi et al. (2010)**: Used an information retrieval approach, breaking images into objects and actions and matching them with predefined sentence templates. Limitations: Struggled with generalization to unseen objects.
2. **Vinyals et al. (2015)**: Introduced the "Show and Tell" model using **CNNs** for feature extraction and **LSTMs** for sentence generation. Strength: Generated more fluent captions. Limitation: Fixed-length vectors led to loss of fine-grained details.
3. **Yang et al. (2019)**: Added **attention mechanisms** to focus on specific parts of the image during caption generation. Strength: Better contextual relevance. Limitation: Struggled with noisy backgrounds and irrelevant regions.
4. **Hodosh et al. (2024)**: Framed captioning as a **ranking problem**, evaluating candidate captions based on semantic relevance. Strength: Enhanced caption precision. Limitation: Relies on high-quality datasets and careful curation of captions.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi crossref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details