Journal Publication of International Research for Engineering and Management (JOIREM)
Volume: 05 Issue: 06 | June-2025

AUTOMATED IMAGE CAPTIONING GENERATOR USING DEEP LEARNING

B. Tanuja[1], CH. Varsha[2], G. Eshwar[3], K. Rohit[4], Mr. K. Vijay[5]

Email: tanujarao777@gmail.com , chikkavarsha09@gmail.com eshwargunji0@gmail.com , 21tq1a6709@siddartha.co.in ,vijaykoraveni.cse@siddartha.co.in

Siddhartha Institute of Technology and Sciences, Narapally, Korremula Road, Ghatkesar, Medchal- Malkajgiri (Dist),500088, Hyderabad, India.

---***---

Abstract - The increasing volume of visual data shared online highlights the need for systems that can understand and describe images automatically. This project presents an automated image captioning generator based on deep learning techniques. The model combines Convolutional Neural Networks (CNNs) for extracting image features with Long Short-Term Memory (LSTM) networks for generating descriptive captions. Trained on a dataset of images paired with human-written captions, the system learns to produce contextually relevant and grammatically correct descriptions. This approach has broad applications in accessibility, digital content management, and image retrieval, bridging the gap between visual understanding and natural language.

Key Words**:** Convolutional Neural Networks (CNNs), Long Short-Term Memory (LSTM), Natural Language Processing (NLP).

1.INTRODUCTION

Visual content is a dominant form of communication in the digital age, yet most images lack meaningful textual descriptions. This limits their accessibility, especially for visually impaired users, and hinders effective indexing and retrieval. Manual annotation is not scalable, creating a need for automated solutions. Image captioning—the task of generating natural language descriptions for images—requires understanding visual elements and their relationships. Recent advances in deep learning, particularly CNNs and LSTMs, have enabled significant progress in this field. This project proposes a deep learning-based system that automatically generates captions for images, contributing to smarter human-computer interaction and improved multimedia accessibility.

In the rapidly evolving field of artificial intelligence, the integration of computer vision and natural language processing has opened new avenues for interpreting and describing visual content. One such advancement is automated image captioning

— the process of generating meaningful textual descriptions for images. This project focuses on building an **automated image captioning system using deep learning techniques**, aiming to bridge the gap between visual data and natural language understanding.

LITERATURE SURVEY

Image captioning has gained significant attention in recent years due to its wide-ranging applications in artificial intelligence, including accessibility tools for visually impaired individuals, content-based image retrieval, social media automation, and human-computer interaction. It represents a challenging problem because it requires the integration of two domains: computer vision for understanding image content and natural language processing (NLP) for generating coherent and contextually relevant descriptions.

Early Approaches

Initial attempts at image captioning involved **template-based** or **retrieval-based** models. Template-based methods used fixed sentence structures filled with detected objects, which led to limited sentence diversity and often lacked contextual relevance. Retrieval-based models, on the other hand, searched for similar images in a dataset and reused their captions, which worked only if visually similar examples were available in the training set.

Deep Learning-based Models

The field saw a major breakthrough with the introduction of **deep learning**, especially **encoder-decoder architectures**, which allowed models to be trained end-to-end. A notable early model was **Show and Tell** by Vinyals et al which used a **Convolutional Neural Network (CNN)**, such as Inception or VGG, to encode images into fixed-length feature vectors, and a **Long Short-Term Memory (LSTM)** network to decode these features into sequences of words. This approach laid the foundation for modern image captioning systems.

Attention Mechanisms

To overcome the limitation of fixed-length encoding, **Xueta introduced the Show, Attend and Tell** model, which integrated **attention mechanisms**. This enabled the model to focus on specific parts of the image while generating each word in the caption, making the model more interpretable and improving performance. Attention mechanisms became a standard component of image captioning models thereafter.

Use of Large Datasets

Datasets play a crucial role in training deep models. Popular datasets like **Flickr8k**, **Flickr30k**, and **MS- COCO** [3] provide thousands of images annotated with multiple human-written captions, allowing models to learn rich visual-semantic relationships. MS-COCO, in particular, includes a diverse set of objects, scenes, and descriptions, making it a benchmark dataset for captioning tasks.

Evaluation Metrics

To evaluate the quality of generated captions, researchers commonly use metrics borrowed from machine translation and text summarization. These include:

BLEU (Bilingual Evaluation Understudy), which measures n-gram overlap between generated and reference captions.

METEOR (Metric for Evaluation of Translation with Explicit Ordering), which considers synonyms and stemming.

CIDE (Consensus-based Image Description Evaluation), which measures consensus between the generated caption and all reference captions using TF-IDF weighting.

Each of these metrics has strengths and weaknesses, and they are often used together for a comprehensive evaluation.

1.1 PROBLEM DEFINITION:

With the vast amount of image data generated daily, there is a growing need for systems that can automatically interpret and describe visual content. Manual annotation is time-consuming and inefficient. Traditional methods lack the ability to understand context and generate meaningful descriptions. This project addresses the challenge of generating accurate image captions using deep learning. By combining Convolutional Neural Networks (CNNs) for image feature extraction and Long Short-Term Memory (LSTM) networks for caption generation, the system aims to produce context-aware, grammatically correct captions that enhance accessibility and improve image-based applications.

1.2 OBJECTIVE OF THE PROJECT:

The main objective of this project is to develop an automated image captioning system that can generate accurate and meaningful textual descriptions of images using deep learning techniques. Specifically, the system aims to:

- Extract visual features from images using Convolutional Neural Networks (CNNs).
- Generate coherent and context-aware captions using Long Short-Term Memory (LSTM) networks.
- Train the model on a large dataset of image-caption pairs to learn semantic relationships.
- Enhance image accessibility for visually impaired users.
- Support applications in image indexing, content retrieval, and digital media management.

1.3 EXISTING SYSTEM

Current image captioning systems primarily use deep learning models that combine computer vision and natural language processing. The widely adopted **encoder- decoder architecture** uses a **Convolutional Neural Network (CNN)** to extract image features and a **Long Short-Term Memory (LSTM)** network to generate captions. A prominent example is the **"Show and Tell"** model, which demonstrated how CNN-LSTM frameworks can be trained end-to-end to produce relevant image descriptions.

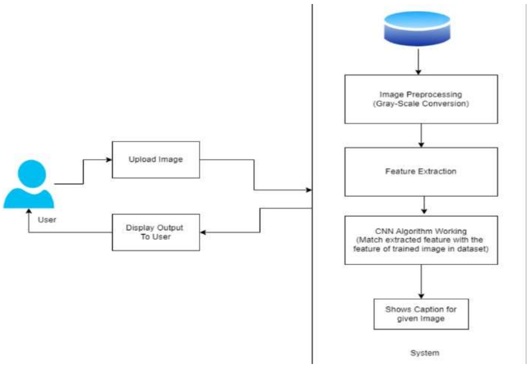
An improvement over this is the **"Show, Attend and Tell"** model, which incorporates **attention mechanisms** to focus on different parts of the image during caption generation, resulting in more accurate and meaningful captions. More recently, **Transformer-based** models like the **Meshed-Memory Transformer** have shown better performance by handling long-range dependencies and improving language fluency.

While these systems have achieved impressive results, they still face limitations in understanding abstract scenes, generating diverse captions, and reducing computational costs.

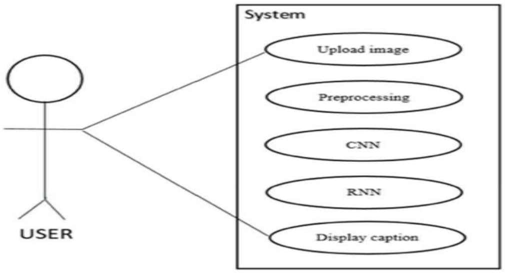
1.4 PROPOSED SYSTEM

The proposed system uses a CNN to extract image features and an LSTM with attention to generate accurate and context-aware captions. By focusing on different image regions during captioning, the attention mechanism improves description quality. Leveraging pre-trained CNN models reduces training time and computational needs. Trained on large datasets like MS-COCO, the system aims to produce diverse, meaningful captions while addressing bias and enhancing semantic understanding for practical applications such as aiding the visually impaired and automated image management.

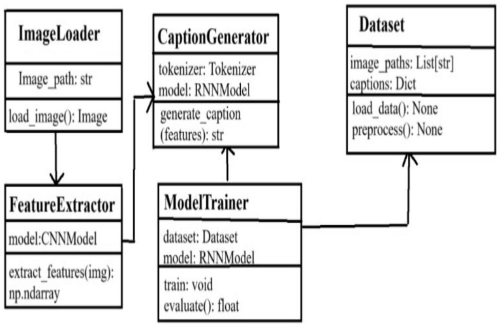
2.5.1 ARCHITECTURAL DIAGRAM


Upload Image
User
Display Output To User
Image Preprocessing (Gray-Scale Conversion)
Feature Extraction
CNN Algorithm Working (Match extracted feature with the feature of trained image in dataset)
Shows Caption for given Image
System

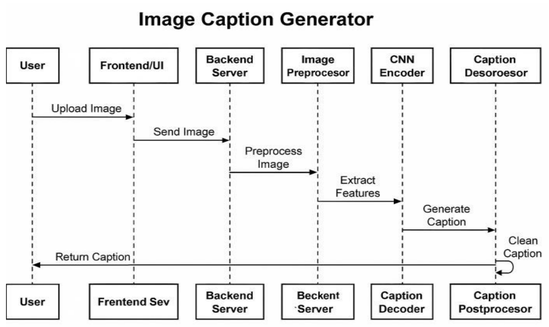
2.5.2 USECASE DIAGRAM



2.5.3 CLASS DIAGRAM



2.5.4 SEQUENCE DIAGRAM



2.6 RESULT

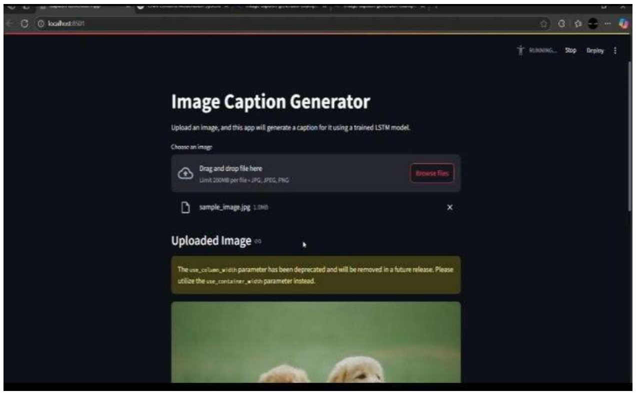


Image Caption Generator

Upload an image, and this app will generate a caption for it using a trained LSTM model.

Choose an image

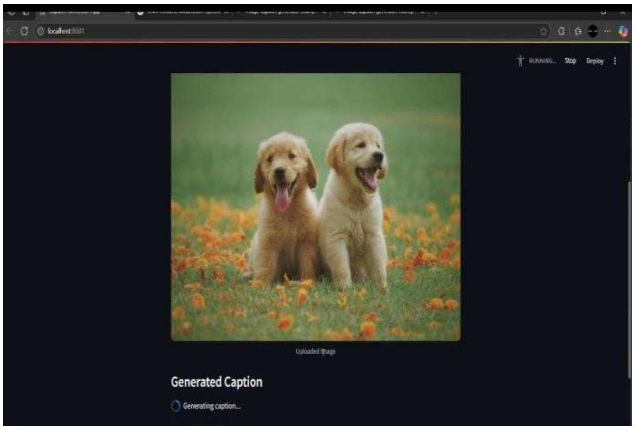
Drag and drop file here
Limit 200MB per file • JPG, JPEG, PNG

sample_image.jpg 1.0MB

Uploaded Image

The use_column_width parameter has been deprecated and will be removed in a future release. Please utilize the use_container_width parameter instead.

Fig -1: Uploading Image

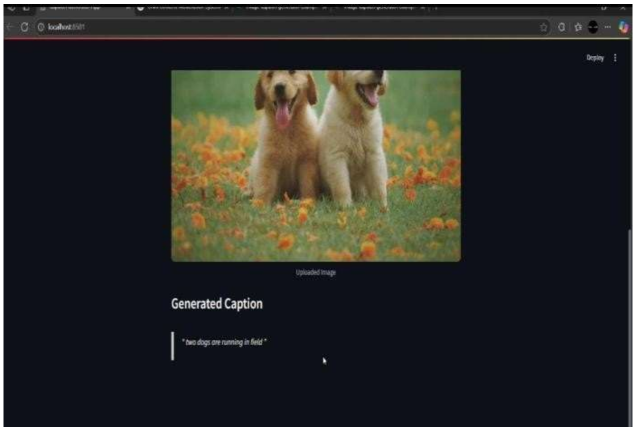


Uploaded Image

Generated Caption

Generating caption...

Fig -2: Generating Caption



Uploaded Image

Generated Caption

" two dogs are running in field "

Fig -3: Generated Caption

3. CONCLUSION

The Image Caption Generator demonstrates the powerful synergy between computer vision and natural language processing by translating visual data into coherent textual descriptions. By leveraging convolutional neural networks (CNNs) for feature extraction and sequence-based models like LSTMs or Transformers for language generation, the system efficiently generates accurate and context-aware captions for a wide variety of images.

Overall, this project provides a strong foundation for real-world applications in areas like digital media, assistive technology, and e-commerce, making visual content more interactive and searchable.

ACKNOWLEDGEMENT

Any endeavor in the field of development is a person's intensive activity. A successful project is a fruitful culmination of efforts by many people, some directly involved and some others who have quietly encouraged and supported.

Salutation to be beloved and highly esteemed institute SIDDHARTHA INSTITUTE OF TECHNOLOGY AND

SCIENCES for grooming us into Computer Science and Engineering graduate, we wish to thank PRINCIPAL DR. M. JANARDHAN for providing a great learning environment.

We wish to express profound gratitude TO MR.K.VIJAY , Assistant Professor and HEAD OF DEPARTMENT, CSE(DATA SCIENCE), for her continuous encouragement to ensure successful results in all my endeavors.

We would like to thank MR.K.VIJAY, Department of CSE, who patiently guided and helped us throughout our project.

We take this opportunity to thank the department's Project Review Co- Ordinator Ms. N. BHARGAVI for all the review meetings, suggestions, and support throughout the project development.

REFERENCES

1. Kumar, R. D., Prudhvi Raj, G., Vijay, K., Kumar, P. S., & Plugmann, P. (2024). Exploring COVID-19 Through Intensive Investigation with Supervised Machine Learning Algorithm. In Handbook of Artificial Intelligence and Wearables (pp. 145-158). CRC Press.
2. Swathi, B., Vijay, K., Sushanth Babu, M., & Dinesh Kumar, R. (2025). Machine Learning Techniques in Cloud Based Intrusion Detection. In The International Conference on Artificial Intelligence and Smart Environment (pp. 557-564). Springer, Cham.
3. Chetan Amritkar and Vaishali Jabade. Image Caption Generation Using Deep Learning Technique. Proceedings- 2018 4th Interna tional Conference on Computing, Communi cation Control and Automation, ICCUBEA 2018, pages 1–4, 2018.
4. Ayan Ghosh, Debarati Dutta, and Tiyasa Moitra. A Neural Network Framework to Generate Caption from Images. Springer Nature Singapore Pte Ltd., pages 171–180, 2020.
5. Yuting Su, Yuqian Li, Ning Xu, and An-An Liu. Hierarchical deep neural network for image captioning. Neural Processing Letters, 52(2):1057–1067, 2020.
6. Yuqing Peng, Chenxi Wang, Yixin Pei, and Yingjun Li. Video captioning with global and local text attention. The Visual Computer, pages 1–12, 2021.
7. Harshitha Katpally and Ajay Bansal. Ensem ble learning on deep neural networks for image caption generation. Proceedings- 14th IEEE International Conference on Semantic Computing, ICSC 2020, pages 61–68, 2020.
8. Muhammad Jaleed Khan and Edward Curry. Neuro-symbolic visual reasoning for multi media event processing: Overview, prospects and challenges. In CIKM (Workshops), 2020.
9. Department of Computer Science Sulabh Katiyar, Samir Kumar Borgohain and Silchar Engineering National Institute of Technology. Comparative evaluation of cnn architectures for image caption generation. International Journal of Advanced Computer Science and Applications, 2021.
10. Automatic Caption Generation for News Images by Yansong Feng, and Mirella Lapata, IEEE (2013).
11. Image Caption Generator Based on Deep Neural Networks by Jianhui Chen, Wenqiang Dong and
12. Show and Tell: A Neural Image Caption Generator by Oriol Vinyal, Alexander Toshev, Samy Bengio, Dumitru Erhan, IEEE (2015).
13. Image2Text: A Multimodal Caption Generator by Chang Liu, Changhu Wang, Fuchun Sun, Yong Rui, ACM (2016).
14. Image Captioning in the Wild: How People Caption Images on Flickr Philipp Blandfort, Tushar Karayil, Damian Borth, Andreas Dengel,German Institute for Artificial Intelligence, Kaiserslautern, Germany.
15. Image Caption Generator Based On Deep Neural Networks Jianhui Chen ,Wenqiang Dong, Minchen Li ,CS Department. ACM 2014.