iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 **Issue:** VIII **Month of publication:** August 2025

DOI: https://doi.org/10.22214/ijraset.2025.73566

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue VIII Aug 2025- Available at www.ijraset.com

Image Caption Generator Using CNN and LSTM Hybrid Models

Ganta Nikhila[1], Dr. M. Dhanalakshmi[2]
[1]*M. Tech, Data Science, Department of Information Technology, Jawaharlal Nehru Technological University Hyderabad, UCESTH, India*
[2]*Professor of IT Dept & Deputy Director of DILT, Department of Information Technology, Jawaharlal Nehru Technological University Hyderabad, UCESTH, India*

Abstract: ***Image captioning refers to the automated process of generating a descriptive sentence that conveys the content of a given image. The developed model receives an image as input and produces an English sentence that accurately represents what is depicted. This area has drawn considerable attention in recent years, particularly in the realm of cognitive computing, due to its reliance on both computer vision and natural language processing techniques. The system utilizes a Convolutional Neural Network (CNN) to analyze and extract visual features from the image, which are then passed to a Long Short- Term Memory (LSTM) network responsible for constructing the descriptive sentence. The CNN functions as the encoder, while the LSTM acts as the decoder. Following caption generation, the model's performance is evaluated to ensure the quality and relevance of the output. This enables the generation of meaningful, human-readable descriptions for various images.***
Keywords: Convolutional Neural Network (CNN), Long Short Term Memory (LSTM), Image Caption.

I. INTRODUCTION

Image captioning combines computer vision and natural language processing to automatically describe the contents of an image in natural language. The project discussed here leverages a deep learning approach, specifically using Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks. The CNN model acts as an encoder to extract high-level visual features, which are then decoded into descriptive text by the LSTM. Using the Flickr8k dataset, the system is trained to map visual elements to meaningful captions, aiming to produce contextually accurate and grammatically sound sentences. This technology has diverse applications, including aiding visually impaired users, enhancing image search, and automating digital content labelling. To improve the quality and accuracy of the generated captions, various training techniques and architectures are employed, such as incorporating attention mechanisms that enable the model to focus on relevant parts of the image during caption generation. This allows for more detailed and specific descriptions, especially in complex scenes. Additionally, the system can be fine-tuned using larger and more diverse datasets, which helps the model learn a broader range of visual.

II. RELATED WORK

The related work on Image Caption Generator using CNN and LSTM Hybrid Models focuses on leveraging the strengths of convolutional neural networks (CNNs) for visual feature extraction and long short-term memory networks (LSTMs) for natural language generation. Previous studies have demonstrated that CNNs such as VGG16, Inception, and ResNet are highly effective in capturing high-level visual representations from images. These features are then passed to LSTM-based models, which are capable of modeling the sequential dependencies in language, making them suitable for generating grammatically and contextually accurate captions. The Show and Tell model by Google, which combined Inception CNN with LSTM, laid the foundation for many subsequent works in this domain. Other improvements include the use of attention mechanisms (e.g., in the Show, Attend and Tell model) to dynamically focus on relevant parts of an image while generating each word. Research has also explored transfer learning, where pre-trained models are fine-tuned for image captioning tasks to reduce training time and improve accuracy. Overall, the integration of CNNs and LSTMs has proven to be a powerful and widely adopted approach for generating human-like descriptions of images.

III. METHODOLOGY

The developed system follows an encoder-decoder framework. The encoder, a pre-trained CNN model (Xception or VGG16), processes each image to produce a fixed-length feature vector summarizing visual content. The decoder, implemented using an LSTM, sequentially generates words to form a caption, conditioned on both the image features and previously generated words.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

The model is trained on the Flickr8k dataset, which provides five captions per image, allowing it to learn diverse ways of describing similar scenes. Preprocessing steps include resizing images to 224×224 pixels, normalization, and text tokenization. Sequence padding ensures uniform input lengths for the LSTM.

An image caption generator is a deep learning-based system designed to produce natural language descriptions for visual inputs. Its architecture typically consists of two primary components: an image encoder that extracts visual features and a language decoder that constructs the corresponding caption.

A. Image Encoder

The image encoder's role is to analyze the input image and extract its most important visual characteristics. This is typically achieved using a pre-trained Convolutional Neural Network (CNN) such as VGG, ResNet, or Inception. These models, having been trained on extensive image datasets, are capable of capturing complex and high-level features. The encoder produces a feature vector as its output, which serves as a compact representation of the image's visual content.

B. Language Decoder

The language decoder generates a descriptive sentence for the input image using the feature vector produced by the image encoder. This component is typically built using a Recurrent Neural Network (RNN), such as a Long Short-Term Memory (LSTM) or Gated Recurrent Unit (GRU) model. The decoder begins with the encoded feature vector as its initial input and generates a word sequence that forms the caption. At each time step, it uses both the previously generated word and the image features to predict the next word.

C. Training

During model training, the image captioning system is adjusted to minimize a loss function that measures the gap between the predicted caption and the reference caption provided by humans. Techniques such as cross-entropy loss or maximum likelihood estimation are commonly employed to assess the similarity between the generated text and the actual caption. This process is carried out using a comprehensive dataset that includes images alongside their corresponding textual descriptions, allowing the model to effectively learn how to associate visual content with meaningful language.

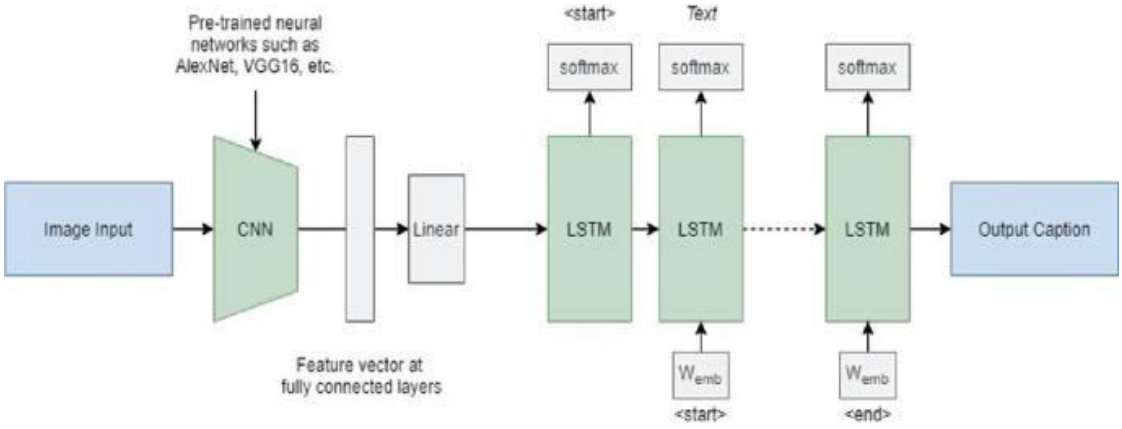


Fig 1. Architecture of Image Caption Generator

D. Modules

1) Image Preprocessing
2) Image based model (CNN)
3) Language based model (LSTM)
4) Caption generation

E. Image Preprocessing

Computers cannot inherently interpret visual content, so the first step involves converting the input image into a standardized pixel matrix of dimensions 224×224×3, where each pixel's RGB color value is mapped accordingly.

After resizing, the image undergoes noise reduction to enhance clarity and remove any unwanted disturbances that may affect feature extraction.

The cleaned image is then transformed into a grayscale format, followed by applying a threshold to distinguish the foreground from the background. Edge detection techniques are applied to identify the boundaries of objects within the image.
The final output of this preprocessing stage is a refined pixel matrix, which serves as the input for the subsequent processing module in the image captioning system.

F. Image Based Model (CNN)

This module uses an enhanced version of a Convolutional Neural Network (CNN), where convolution and pooling layers function as the primary mechanisms for extracting meaningful visual features.
The input to this module is the processed pixel matrix obtained from the image preprocessing stage.
It identifies and extracts key features from the input matrix and compiles them into a feature vector.
The extracted details include object identities, actions or movements associated with those objects (verbs), object colors, and most importantly, the spatial or contextual relationships between objects within the image.
The generated feature vector is subsequently forwarded to the next stage of the processing pipeline. Before doing so, the vector is resized or linearly transformed to match the required input dimensions of the LSTM network used in the caption generation phase.

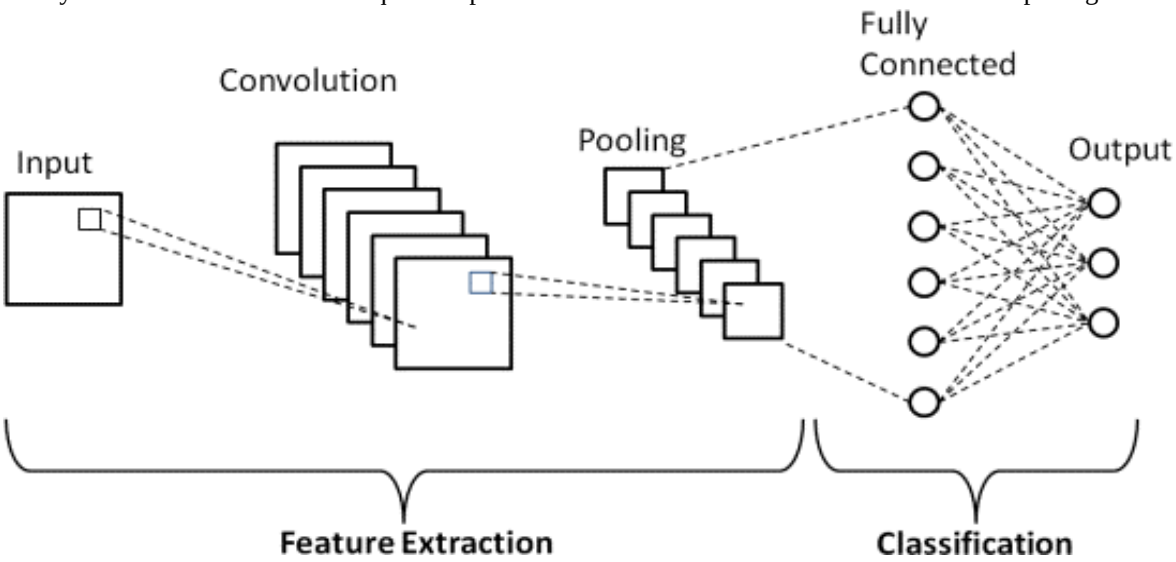


Fig Architecture of Convolutional Neural Network

IV. LANGUAGE BASED MODEL(LSTM)

In the "Image Caption Generator Using CNN and LSTM Hybrid Models" project, the language-based model is implemented using a Long Short-Term Memory (LSTM) network, which serves as the core component for generating captions in natural language. After the image features are extracted by the Convolutional Neural Network (CNN), these features are fed into the LSTM model, which functions as a sequence generator. The LSTM takes these visual features along with the previously generated words (or ground truth words during training) to predict the next word in the caption sequence. To process textual input, words are first converted into dense vector representations using word embeddings, enabling the model to understand the semantic relationships between words. The LSTM is trained using a method called teacher forcing, where the correct word at each time step is provided as input during training. During inference, the model uses techniques like greedy search or beam search to generate coherent and contextually relevant captions word by word, starting with a special start token and ending when a stop token is predicted. Overall, the LSTM-based language model enables the system to translate visual content into descriptive natural language effectively.

G. Caption Generator

The final stage in the system is the Caption Generation module, which receives input from the preceding Language-Based Module.
The primary objective of this component is to produce a complete and coherent caption in a linear word sequence, based on the information processed by the LSTM network.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

V. RESULTS AND EVALUATIONS

```
In [31]: generate_caption("1002674143_1b742ab4b8.jpg")

----------------------Actual----------------------
startseq little girl covered in paint sits in front of painted rainbow with her hands in bowl endseq
startseq little girl is sitting in front of large painted rainbow endseq
startseq small girl in the grass plays with fingerpaints in front of white canvas with rainbow on it endseq
startseq there is girl with pigtails sitting in front of rainbow painting endseq
startseq young girl with pigtails painting outside in the grass endseq
--------------------Predicted--------------------
startseq little girl in pink dress is lying on the side of the grass endseq
```



Fig Image Caption 1

```
generate_caption("1028205764_7e8df9a2ea.jpg")

----------------------Actual----------------------
startseq man and baby are in yellow kayak on water endseq
startseq man and little boy in blue life jackets are rowing yellow canoe endseq
startseq man and child kayak through gentle waters endseq
startseq man and young boy ride in yellow kayak endseq
startseq man and child in yellow kayak endseq
--------------------Predicted--------------------
startseq man in yellow kayak in the water endseq
```



Fig Image Caption 2

```
In [32]: generate_caption("101669240_b2d3e7f17b.jpg")

----------------------Actual----------------------
startseq man in hat is displaying pictures next to skier in blue hat endseq
startseq man skis past another man displaying paintings in the snow endseq
startseq person wearing skis looking at framed pictures set up in the snow endseq
startseq skier looks at framed pictures in the snow next to trees endseq
startseq man on skis looking at artwork for sale in the snow endseq
--------------------Predicted--------------------
startseq two people are hiking up snowy mountain endseq
```



Fig Image Caption 3

```
generate_caption("1009434119_febe49276a.jpg")

----------------------Actual----------------------
startseq black and white dog is running in grassy garden surrounded by white fence endseq
startseq black and white dog is running through the grass endseq
startseq boston terrier is running in the grass endseq
startseq boston terrier is running on lush green grass in front of white fence endseq
startseq dog runs on the green grass near wooden fence endseq
--------------------Predicted--------------------
startseq black and white dog is running through the grass endseq
```



Fig Image Caption 4

1) Performance Analysis of Image Captioning Models

```
from nltk.translate.bleu_score import corpus_bleu
actual, predicted = list(), list()
for key in tqdm(test):
    # Ground truth captions
    captions = mapping[key]
    # Predict caption for the image
    image = features[key][0]  # Extract the feature vector for this image
    y_pred = predict_caption(model, image, tokenizer, max_length)
    # Tokenize captions for BLEU score
    actual_captions = [caption.split() for caption in captions]
    predicted_caption = y_pred.split()
    actual.append(actual_captions)
    predicted.append(predicted_caption)
# Compute BLEU scores
print("BLEU-1: %f" % corpus_bleu(actual, predicted, weights=(1.0, 0, 0, 0)))
print("BLEU-2: %f" % corpus_bleu(actual, predicted, weights=(0.5, 0.5, 0, 0)))
print("BLEU-3: %f" % corpus_bleu(actual, predicted, weights=(0.33, 0.33, 0.33, 0)))
print("BLEU-4: %f" % corpus_bleu(actual, predicted, weights=(0.25, 0.25, 0.25, 0.25)))
```

```
100% 810/810 [10:08<00:00,  1.31it/s]

BLEU-1:0.5044
BLEU-2:0.5012
BLEU-3:0.4834
BLEU-4:0.4003
```

BLEU Score is used to evaluate the predicted text against a reference text, in a list of tokens.The reference text contains all the words appended from the captions data (actual_captions)A BLEU Score more than **0.4 is considered a good result**, for a better score increase the no. of epochs accordingly.

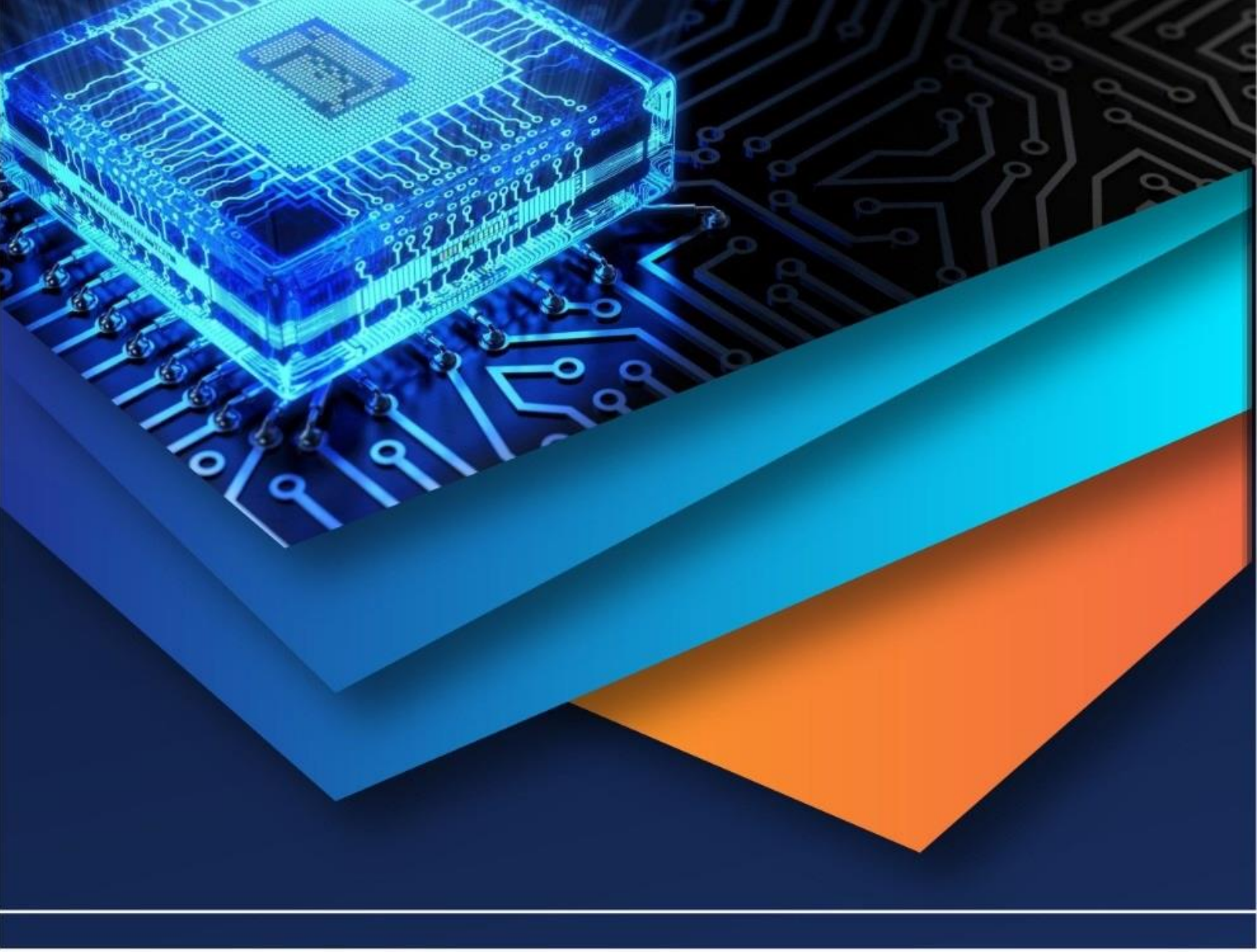
VI. CONCLUSION

The Image Caption Generator project successfully showcases the combination of computer vision and natural language processing to produce meaningful textual descriptions of images. It utilizes a pre-trained VGG16 CNN model to extract rich visual features, which are then fed into an LSTM-based sequence model that constructs captions one word at a time.The application of transfer learning enhances the system's efficiency by reducing training duration and improving accuracy. Additionally, techniques such as tokenization and sequence padding ensure that the text data is structured appropriately for training the language model.Model performance is assessed through both qualitative analysis (by reviewing generated captions) and quantitative evaluation using BLEU scores, confirming that the outputs are contextually accurate and grammatically sound.This project emphasizes the effectiveness of deep learning in addressing complex, multimodal challenges and lays the groundwork for practical implementations like automated image tagging, assistive technologies for the visually impaired, and intelligent media organization tools.

REFERENCES

[1] Base Paper: Katiyar, S., &Borgohain, S. K. (2021). Comparative evaluation of CNN architectures for image caption generation. arXiv preprint arXiv:2102.11506.
[2] Kalena, P., Malde, N., Nair, A., Parkar, S., & Sharma, G. (2019). Visual Image Caption Generator Using Deep Learning. In 2nd International Conference on Advances in Science & Technology.
[3] Xu, K., CA, U., Ba, J. L., CA, U., Kiros, R., EDU, T., &Bengio, Y. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention (Supplementary Material).
[4] Jia, X., Gavves, E., Fernando, B., &Tuytelaars, T. (2015). Guiding the long-short term memory model for image caption generation. In Proceedings of the IEEE international conference on computer vision (pp. 2407-2415).
[5] Yagcioglu, S., Erdem, E., Erdem, A., &Cakıcı, R. A Distributed Representation Based Query Expansion Approach for Image Captioning (Supplementary Material).
[6] Kinghorn, P., Zhang, L., & Shao, L. (2018). A region-based image caption generator with refined descriptions. Neurocomputing, 272, 416-424.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)