1.INTRODUCTION

1.1 Problem Statement:

Image captioning presents a significant challenge in Artificial Intelligence, requiring systems to understand images and generate grammatically correct descriptions. The existing methods facing issues in accurately describing images with proper context. To address this, a hybrid system Convolutional Neural Networks (CNN) for feature extraction and Long Short-Term Memory (LSTM) networks for caption generation is proposed. However, the existing system struggling to capture suitable image differences and contextual nuances. Therefore, there is a need for an improved image captioning system that effectively integrates CNN and LSTM models to accurately describe images with contextual relevance. This research aims to develop such a system, utilizing the Flickr8K dataset for training and evaluation. The system efficiency will increased by generating suitable captions for given images.BLEU Score is used as a metric to evaluate the performance of the trained model.

1.2 Objective of project:

The objective of the project is to predict the captions for the input image. The dataset consists of 8k images and 5 captions for each image. The features are extracted from both the image and the text captions for input. The features will be concatenated to predict the next word of the caption. CNN is used for image and LSTM is used for text. BLEU Score is used as a metric to evaluate the performance of the trained model.

1.3 Scope & Limitations of the project:

- Lack of understanding context.
- Limited Dataset coverage
- Difficulty with Ambiguity
- Inability to Incorporate External Knowledge

2. LITERATURE SURVEY

- [1] One of the research study presents a model employing pre-trained deep learning, specifically the VGG model, for generating captions from images. By comparing the model's output with human-provided captions, it achieves an accuracy of around 75%, indicating a close resemblance between generated and human captions. This approach enhances the capability of automated caption generation, bridging the gap between machine-generated and human-provided descriptions of images.
- [2] The image caption generator in few studies utilizes the Flickr_8k database, comprising 8000 diverse images, each with five captions. The dataset is divided into 6000 training, 1000 validation, and 1000 testing images. Through rigorous training and testing, the model effectively generates accurate captions. It employs a combination of Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN), where CNN acts as the encoder and RNN as the decoder, ensuring grammatically correct captions with appropriate labels.
- [3] In an image caption generator, the VGG16 model serves as a sophisticated filter for images, identifying crucial features such as shapes and objects. These features form a summary of the picture, which is then utilized by another component of the system to generate a descriptive sentence. VGG16 essentially aids in comprehending the content of the image, allowing the caption generator to translate this understanding into words.
- [4] Some studies presents a framework for generating descriptive captions from images, utilizing the Flickr8K dataset containing 8000 images, each paired with five descriptions. The model employs a neural network to automatically analyze images and generate English captions. The generated captions are categorized into error-free descriptions, descriptions with minimal errors, somewhat related descriptions, and unrelated descriptions, showcasing the system's ability to produce diverse outputs based on image content.

- [5] The integration of VGG16, LSTM, and CNN in image caption generation is significant for enhanced performance. VGG16 acts as a robust feature extractor, capturing detailed information from images. LSTM complements VGG16 by generating coherent captions, leveraging its sequential data understanding. The synergy between CNN and LSTM addresses challenges in combining visual and linguistic information, leading to a comprehensive image understanding. Ongoing research may unveil novel approaches, refining the synergy between these components for improved image captioning.

3. PROPOSED METHODOLOGY

3.1 Existing System

The complete system is a combination of three models which optimizes the whole procedure of caption description from an image. The models are:

(a)Feature Extraction Model :

The model uses a VGG16 architecture to efficiently extract the features from the images using a combination of multiple 3*3 convolution layers. The output of a VGG16 network would be vectors of size 1*4096, which are used to represent the features of the images.

(b) Encoder Model :

The encoder model, is primarily responsible for processing the captions of each image fed while training. The output of the encoder model is again vectors of size 1*256 which would again be an input to the decoder sequences Initially the captions present with each images are tokenized ie the words in the sentences are converted to integers so that the neural network can process them efficiently.

(c)Decoder model:

The decoder model, is basically the model which concatenates both the feature extraction model and encoder model and produces the required output which is the predicted word given an image and the sentence generated till that point of time.

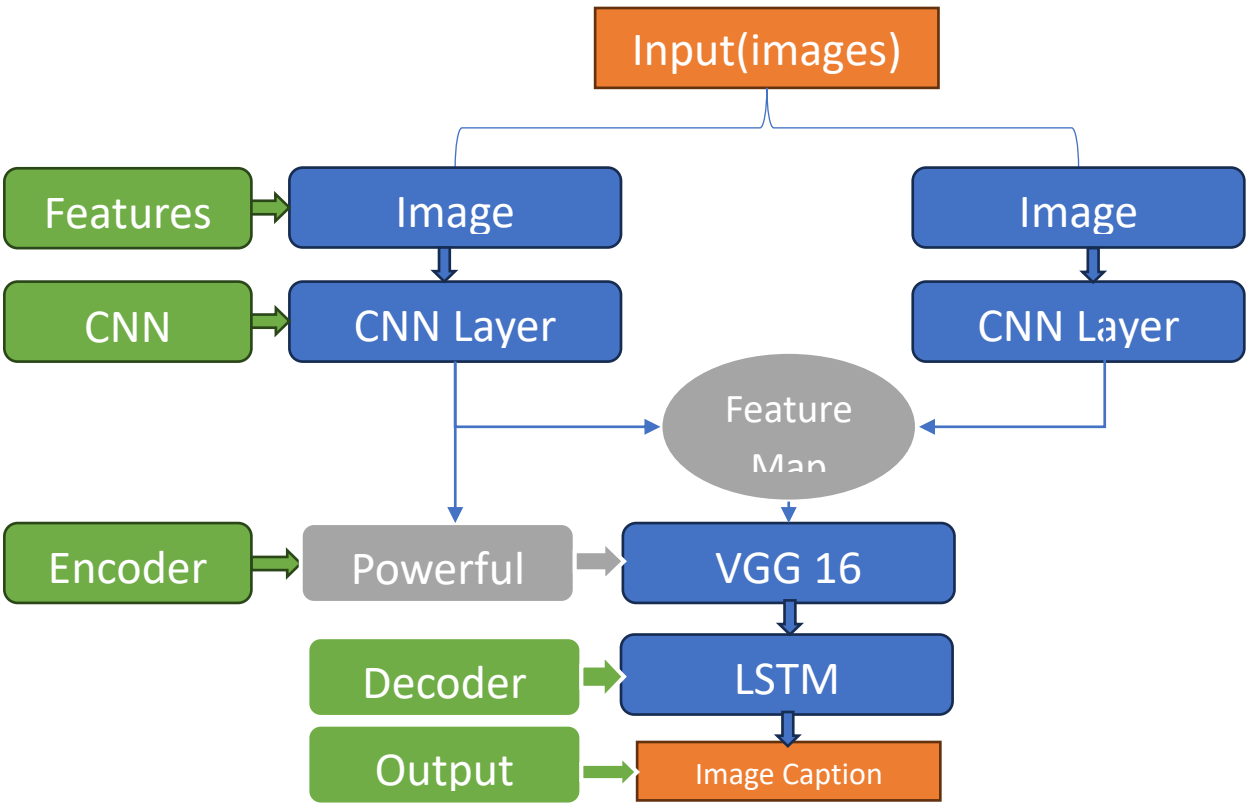
3.2 Proposed System

- Our model is not perfect and may generate incorrect captions sometimes. In the next phase, we will be developing models which will use Inceptionv3 instead of VGG as the feature extractor. Then we will be comparing the 4 models thus obtained i.e. VGG+GRU, VGG+LSTM, Inceptionv3+GRU, and Inceptionv3+LSTM . This will further help us analyze the influence of the CNN component over the entire network.
- Our model is trained on the Flickr 8K dataset which is relatively small with less variety of images. We will be training our model on the Flickr30K and MSCOCO datasets which

will help us to make better predictions. Other optimizations include tweaking the hyperparameters like batch size, number of epochs, learning rate etc and understanding the effect of each one of them on our model.

- Evaluation metrics such as BLEU score validate the model's performance. Once validated, the system is tested on unseen data and can be deployed for real-time caption generation. Future directions include exploring advanced techniques like attention mechanisms for further improvement. Overall, the proposed system offers a robust solution for generating accurate and contextually rich captions for images.

3.3 Architecture



- The CNN receives an image as input.
- The CNN processes the image through multiple convolutional layers. Each layer uses filters to identify specific features in the image, like edges or shapes. The filters slide across the image, and their activations are stored in a feature map.
- As the image goes through more convolutional layers, the feature maps become more complex, encoding increasingly intricate information about the image.
- Finally, the CNN can use the learned features to perform a task, such as reconstructing the original image or recognizing objects within it.

3.4 Flowchart

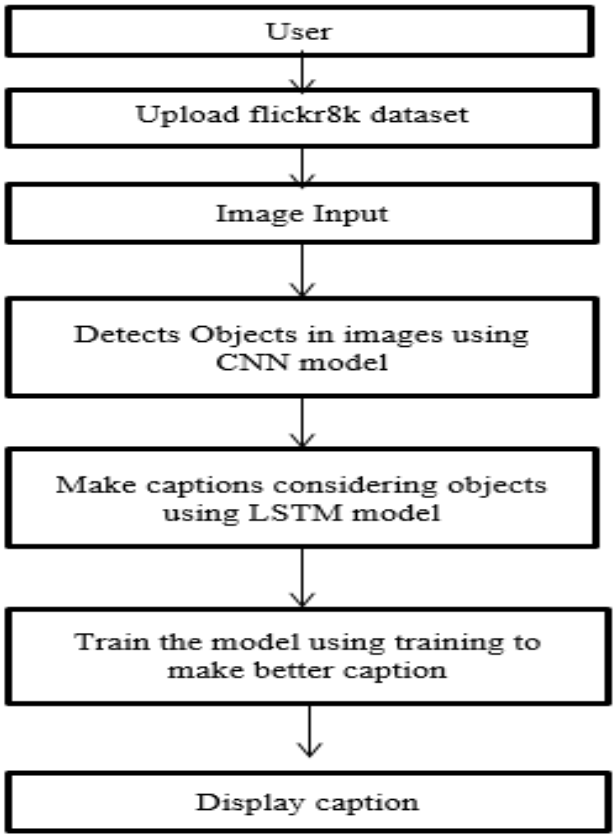


Figure: Flow Chart

- User uploads an image: The process starts with a user uploading an image to the system.
- Image goes through a CNN model: The CNN model analyzes the image to identify objects and their locations within the image.
- Captions are generated using the LSTM model: The LSTM model uses the information about the objects detected by the CNN model to generate a caption that describes the image.
- Model is trained to improve captions: The system can be trained on a dataset of images and captions to improve the accuracy of the captions generated by the LSTM model.
- Caption is displayed: Finally, the generated caption is displayed for the user.

3.5 Methods & Algorithms

3.5.1 Convolutional Neural Network (CNN):

Purpose: CNNs are primarily used for image analysis and feature extraction.

Functionality:

- Multilayer Architecture: CNNs consist of multiple layers (convolutional, pooling, and fully connected) that learn hierarchical features from raw pixel data.
- Convolutional Layers: These layers apply convolutional filters to extract local patterns (edges, textures, shapes) from the input image.
- Pooling Layers: Reduce spatial dimensions while preserving important features.
- Fully Connected Layers: Process the extracted features and make predictions.
- Significance: CNNs can automatically learn relevant features from images, making them effective for tasks like object recognition and localization.

3.5.2Long Short-Term Memory (LSTM):

- Purpose: LSTMs are designed to handle sequential data, such as natural language.
- Functionality:
- Recurrent Architecture: LSTMs have recurrent connections that allow them to maintain memory over time steps.
- Cell State and Gates: LSTMs use a cell state to store information and three gates (input, forget, and output) to control the flow of information.
- Avoiding Vanishing Gradient Problem: LSTMs mitigate the vanishing gradient problem by allowing gradients to flow through time steps.
- Significance: LSTMs capture long-term dependencies and context, making them suitable for tasks like language modeling and sequence-to-sequence tasks.

3.5.3 Encoder-Decoder Model:

Purpose: Combines an encoder (to process input data) and a decoder (to generate output sequences).

Functionality:

1)Encoder (VGG16):

Uses a pre-trained CNN (e.g., VGG16) to encode input images into a fixed-length feature vector.

The feature vector captures high-level visual information.

2)Decoder (LSTM):

Takes the encoded feature vector and generates captions word by word.

Learns to predict the next word based on context and previously generated words.

Significance: Encoder-decoder models bridge the gap between visual and textual information, enabling image captioning.

3.5.4 Flickr8K Dataset:

Purpose: Used for training and evaluating the image captioning model.

- Content:

  Contains 8,000 images along with corresponding human-generated captions.

  Captions provide context and describe the content of each image.

- Training Process:

  The model learns to associate image features (from VGG16) with relevant captions.

  It generalizes to generate contextually relevant captions for unseen images.