Project Title: “Image Captioning Using CNN and LSTM”

Abstract:

With advancements in artificial intelligence, generating meaningful captions for images has become an important challenge in computer vision and natural language processing. This project aims to develop an image captioning system that automatically generates descriptive sentences for images by leveraging deep learning techniques. The system integrates Convolutional Neural Networks (CNNs) for feature extraction from images and Long Short-Term Memory (LSTM) networks for sequential text generation.

The methodology involves training a deep learning model on a dataset containing images and corresponding captions. Pre-trained CNN models such as InceptionV3 or ResNet are used to extract image features, which are then passed to an LSTM-based language model for caption generation. TensorFlow and Keras are used for model development, with data preprocessing techniques such as tokenization, padding, and embedding applied to improve text processing.

Initial results demonstrate that combining CNNs and LSTMs effectively captures the visual and textual relationships in images, producing high-quality captions. The model achieves strong performance in evaluation metrics such as BLEU and METEOR scores, indicating its applicability for real-world applications in automated image annotation, accessibility tools, and content generation.

Step-wise Solution Approach:

Step 1: Data Collection – Gather a dataset of images with corresponding captions (e.g., MS COCO dataset).

Step 2: Data Preprocessing – Perform text tokenization, word embedding, padding, and image normalization to prepare the dataset.

Step 3: Model Selection – Use a pre-trained CNN model (InceptionV3 or ResNet) for image feature extraction and an LSTM model for caption generation.

Step 4: Training – Train the model using TensorFlow and Keras, optimizing hyperparameters to improve caption accuracy.

Step 5: Inference and Real-Time Implementation – Implement a pipeline for generating captions for new images in real-time.

Step 6: Performance Evaluation – Assess model accuracy using BLEU, METEOR, and CIDEr scores to measure caption quality.