**ARTIFICAL INTELLIGENCE AND MACHINE LEARNING (AIML) – PROJECT**

NAMES: SEC:5

Anubhav Joshi-2320030416  
Varun reddy-2320030047  
**Problem Statement:** Design an algorithm to automatically generate captions for images based on their content.  
The goal is to design an algorithm that can automatically generate meaningful and accurate captions for images based on their content. This involves analyzing the image to detect objects, actions, and contextual relationships, and then converting this information into coherent sentences. The system should be able to generate captions that describe the overall scene and capture both specific objects and their interactions.

**Dataset:** To build such a system, the following types of datasets are commonly used:

1.**MS COCO (Microsoft Common Objects in Context) :** A popular dataset containing over 330,000 images, each with five human-annotated captions describing the scene.

**2.Flickr8k / Flickr30k:** Datasets with 8,000 and 30,000 images respectively, each paired with five captions.

**3.Visual Genome:** A dataset that provides detailed scene descriptions, object annotations, attributes, and relationships between objects in images.  
These datasets contain a variety of objects, scenes, and activities with corresponding human-generated captions, which are essential for training the model.  
  
 **ALOGRITHM:**

**Preprocessing and Feature Extraction:**

**1.Image Input :** Feed the input image into a pre-trained convolutional neural network (CNN), like ResNet or Inception, to extract a feature vector representing the image content**.**

**2.Text Input :** Preprocess the text data (captions) by tokenizing the sentences and converting them into sequences of word embeddings (using Glove, Word2Vec, or pre-trained embeddings from models like BERT).

**Model Architecture:**

**1.Encoder:** Use a CNN (e.g., ResNet-50 or InceptionV3) as the encoder to extract visual features from the image.

**2.Decoder:** Use a Recurrent Neural Network (RNN) with Long Short-Term Memory (LSTM) units to decode the image features into sentences. The LSTM generates one word at a time based on the previously generated words and the image features.

The model follows an \*\*encoder-decoder\*\* architecture, where:

- The encoder (CNN) extracts high-level image features.

- The \*\*decoder(RNN/LSTM) generates a caption word by word.

**3.Training:**During training, use the image as input and the associated caption as the target output. The model will learn to generate captions by minimizing the loss function, typically cross-entropy, between the predicted word and the actual word at each time step.

The algorithm is trained to predict the next word in the sequence based on the image features and the previous words.  
  
 **4.Beam Search for Caption Generation:**

To generate captions during inference, use beam search instead of greedy decoding. Beam search explores multiple possible caption sequences at each step and selects the sequence with the highest overall probability**.  
  
5. Evaluation:**

- Evaluate the model using metrics such as BLEU, METEOR, ROUGE, and CIDEr, which measure the similarity between the generated captions and the ground-truth captions.

**EXPECTED OUTPUT:**

The model should be able to generate captions for unseen images. For example:

**1. Input:**

[Sample Image: A man riding a bicycle on a city street]

**2. Output:**

- "A man is riding a bicycle down a busy street."

- "A person cycles on a road with cars in the background."

**3.Performance Metrics:**

**BLEU Score:** Measures how well the generated captions match the human-provided captions in the dataset.

**METEOR / CIDEr:** Captures the semantic similarity between generated and actual captions.

In the end, the system will be capable of generating natural-language descriptions of images that convey essential aspects of the scene, such as objects, actions, and interactions.