**Caption Generator Using Computer Vision**

**Abstract**   
**Project Objectives:** The primary objective of this project is to develop an automated system that can generate meaningful captions for images by leveraging deep learning techniques. This system will combine computer vision and natural language processing to describe image content accurately.

**Methodology:** The project utilizes a two-stage approach: feature extraction and sequence generation. A pre-trained Convolutional Neural Network (CNN) such as InceptionV3 extracts visual features from input images. These features are then fed into a Long Short-Term Memory (LSTM) network, which generates textual descriptions based on training data. The model is trained on a dataset like MS-COCO, where image-caption pairs are used to learn the mapping between images and textual descriptions.

**Key Findings:** The developed system demonstrates the ability to generate meaningful and contextually relevant captions for a variety of images. By using pre-trained deep learning models, the accuracy of caption generation improves significantly. The integration of modern transformer-based models, such as CLIP, further enhances captioning capabilities.

**Step-wise Solution Approach:**

* **Step 1:** Data Collection and Preprocessing – Gathering and cleaning the dataset, tokenizing captions, and preparing images.
* **Step 2:** Feature Extraction – Using CNN models like InceptionV3 to extract image feature vectors.
* **Step 3:** Caption Generation Model – Training an LSTM-based sequence generator using extracted features and textual captions.
* **Step 4:** Model Optimization – Fine-tuning hyperparameters and testing performance.
* **Step 5:** Evaluation – Comparing generated captions with ground truth using BLEU and METEOR scores.
* **Step 6:** Deployment – Integrating the trained model into an application for real-world use.

**Team Members –**

* Abhik Rajgaria – E22CSEU1357
* Meenal Khatri – E22CSEU1355
* Divyansh Lal – E22CSEU1237