**Caption Generator for Bipedal Humanoid Using Computer Vision**

**Abstract:** This project aims to integrate an automated image captioning system into a bipedal humanoid robot. By leveraging deep learning techniques, the humanoid will generate meaningful descriptions of its environment based on visual inputs. The system combines computer vision and natural language processing (NLP) to provide real-time captions, enhancing situational awareness and interaction capabilities.

**Project Objectives:**

* Develop a vision-based caption generator for the humanoid robot.
* Enable the humanoid to describe its surroundings using deep learning techniques.
* Improve human-robot interaction by providing contextual descriptions of the environment.
* Optimize the model for real-time processing on embedded hardware.

**Methodology:** The system follows a two-stage approach: feature extraction and sequence generation.

1. **Feature Extraction:** A pre-trained Convolutional Neural Network (CNN) such as InceptionV3 or MobileNet extracts visual features from the humanoid’s camera feed.
2. **Caption Generation:** The extracted features are fed into a Long Short-Term Memory (LSTM) network, trained on an image-caption dataset like MS-COCO, to generate descriptive text.
3. **Integration with Speech Synthesis:** The generated captions are converted into speech using a Text-to-Speech (TTS) engine to enable auditory feedback.

**Key Findings:**

* The system successfully generates contextually relevant captions in real time.
* Using lightweight CNN models enhances processing speed for embedded hardware.
* Transformer-based models such as CLIP can further improve captioning accuracy.
* Integration with speech synthesis improves interaction capabilities.

**Step-wise Solution Approach:**

1. **Data Collection and Preprocessing:**
   * Gather and preprocess image-caption datasets.
   * Tokenize captions and resize images for compatibility with CNN models.
2. **Feature Extraction:**
   * Use CNN models like InceptionV3 or MobileNet to extract feature vectors from real-time camera input.
3. **Caption Generation Model:**
   * Train an LSTM-based model on image-caption datasets.
   * Explore transformer-based alternatives for improved accuracy.
4. **Model Optimization:**
   * Fine-tune hyperparameters to balance accuracy and speed.
   * Optimize the model for deployment on embedded systems like Raspberry Pi or Jetson Nano.
5. **Evaluation:**
   * Compare generated captions with ground truth using BLEU and METEOR scores.
   * Conduct real-world tests in dynamic environments.
6. **Deployment in Humanoid:**
   * Integrate the trained model with the humanoid's vision module.
   * Implement a TTS engine for verbalizing captions.
   * Test and refine real-time performance.

**Team Members:**

* Abhik Rajgaria – E22CSEU1357
* Divyansh Lal – E22CSEU1237
* Smriddhi Parashar – E22CSEU1524