**Title - Image Captioning Using Deep Learning and Neural Translation**

**Abstract**

This project focuses on developing an AI system that analyzes an image, generates a meaningful caption or description, and translates it into different languages. The goal is to create a user-friendly tool that automates image understanding and multilingual accessibility. The system has three key components: an **image captioning model**, a **translation module**, and an **evaluation metric** to assess translation accuracy.

For caption generation, we use **YOLOv4 and YOLOv8**, which are deep learning models for object and action recognition. YOLOv4 detects objects in the image, while YOLOv8 identifies actions to improve caption quality. The detected objects and actions are processed by an **NLP-based text generation model** to create meaningful captions. The generated captions are then translated into multiple languages using **MarianMT**, a neural machine translation model.

To measure translation accuracy, we apply **BLEU (Bilingual Evaluation Understudy) scores** and **cosine similarity**, which compare original and translated captions. This project has applications in **education, accessibility, and global communication**, ensuring that AI-generated captions are accurate, meaningful, and easily translatable across languages.

**Project Objectives**

* Develop an AI system that generates **captions/descriptions** from images.
* Translate the generated captions into multiple languages.
* Evaluate translation accuracy to reduce meaning loss.
* Create a simple and interactive tool for users.

**Methodology**

Our system follows a structured approach to generate captions from images, translate them, and evaluate translation accuracy. We use **YOLOv4 and YOLOv8** for object and action recognition. YOLOv4 detects objects, while YOLOv8 identifies associated actions, improving caption quality. The detected objects and actions are then processed using an **NLP model** to generate meaningful captions.

For translation, we use **MarianMT**, a neural machine translation model, to convert captions into multiple languages. To assess translation accuracy, we apply **BLEU scores** and **cosine similarity metrics**, comparing original and translated captions. BLEU evaluates word overlap, while cosine similarity checks semantic closeness.

By integrating **action recognition**, we ensure more **detailed and context-aware captions**, following the research paper’s approach. The system allows users to upload images and receive captions in multiple languages, making it useful for accessibility, education, and global communication.

**Key Findings**

* AI models like YOLOv4 and YOLOv8 improve image captioning accuracy.
* Action recognition helps generate **more meaningful captions** than object detection alone.
* Some languages experience **higher translation loss** due to linguistic differences.
* AI-based translation methods **improve multilingual accessibility** in captioning.

**Stepwise Solution Approach**

Step 1: Implement YOLOv4 for object detection and YOLOv8 for action recognition.  
Step 2: Use an NLP model to generate meaningful captions.  
Step 3: Translate captions into multiple languages using MarianMT.  
Step 4: Measure translation accuracy using BLEU scores and cosine similarity.  
Step 5: Develop an interactive interface for user image uploads and caption generation.

**Reference**

[1] Charles Rowles-Davies, Georgios Mastorakis, "From Images to Storytelling: Using Deep Learning and Natural Language Processing Modules to Compile Short Stories," 2024.

**Team**

1. Lakshya Deewan (E22CSEU1012)
2. Rajat Mani Bhardwaj (E22CSEU1013)