TEAM – 2 REPORT

Project Track 1a: Set up and generate rare-human actions with AI video generator

SE21UARI038 – KARTHEEK GOALLA

SE21UARI040 – SREE CHAKRITHA

SE21UARI041- HANGARGA AKSHAY KUMAR

SE21UARI042 – HARIVAMSI PODURI

OVERVIEW:

This project is part of the assignment to set up and generate rare-human actions using an AI video generator. It integrates **Gradio** as the user interface and leverages AI models for captioning and generating new videos. Below is a comprehensive breakdown of the work, steps, and outputs.

OBJECTIVE:

1. **Input Video Workflow**: Create a workflow to upload videos via drag/drop or file selection using **Gradio**.
2. **Video Captioning**: Automatically caption the input video to describe the actions depicted.
3. **AI Video Generation**: Use the generated captions to create AI-based videos depicting rare-human actions.
4. **Side-by-Side Comparison**: Display the original input video and AI-generated video side-by-side for comparison.
5. **Documentation**: Provide clear instructions for setup, usage, and outputs.
6. **Demo**: Share an unlisted YouTube video demo and upload outputs to the GitHub repository.

➔ **Video captioning:**

This Python script creates a Gradio-based interface for generating captions from videos using a Vision-to-Language model. Below is a detailed explanation of the code, including required libraries, their purpose, and the script's workflow.

- **Dependencies and Installation:**
  pip install torch transformers gradio av numpy
- **Device Setup**: The script checks if a GPU is available. If not, it defaults to using the CPU for model inference:
  device = "cuda" if torch.cuda.is_available() else "cpu"
- Run the script in an environment (e.g., Google Colab, Jupyter Notebook, or a local Python environment).
- The Gradio app will launch, displaying a web interface with a video uploader and a caption box.

- Upload a video, and the model will generate a caption for it.

  - We have tried 3 models on a whole
  - 1) blip2 model
  - 2)actbert
  - 3)space time GPT
- Although we didn’t get any satisfactory results but space time gpt was better at generating captions.

➔ **Video generating:**

This Python script generates videos from textual captions using the LattePipeline. The pipeline encodes textual descriptions (captions) into embeddings and uses these embeddings to synthesize video frames. Below is a detailed report on the dependencies, installation, workflow, and explanations for each section of the code.

- **Dependencies and Installation**
  pip install bitsandbytes torch diffusers transformers pandas imageio torchvision
- **Workflow**
  - setup garbage collection (Clears memory to prevent GPU or CPU from running out of space during video generation.)
  - Configuration
  - loading video captions
  - initialize text encoder
  - initialize latte pipeline
  - process captions and generate embedding
  - generate video

- Reads video names and captions from a CSV file.
- Encodes captions into embeddings using a quantized text encoder.
- Generates videos using these embeddings with the LattePipeline.
- Saves the videos in Google Drive.

  Note : Ensure your environment has enough GPU memory, especially for generating high-quality videos.

- We have tried several model in which we were able to achieve few satisfactory results but not promising compared to text 2 video model