Multimodal LLM

# Project Objectives

1. Developing a system that can automatically summarize the content of videos and images containing text or subtitles using LLMS.
2. Analyse and interpret the key differences between videos and image sequences, focusing on:
   1. Motion and actions within the video
   2. Sequence and order of information presented in images and video stream.

# Approaching the 1st Objective

We approach the problem with a solution that utilizes different models for each part of the problem. The idea is to perform video analysis using a separated model, audio analysis, using a separate model, and then combining them with a final LLM to get an overall description. Following are the parts and their proposed solutions (might need tweaking depending on what works, and what doesn’t) –

## Video Analysis

Scouted for 2 models (Looking for more, and potentially modifying one of these to make them lighter on computation by trading in some functionality)

**PLLaVA (Pooling LLaVA)** – PLLaVA adapts image Multimodal Large Language Models (MLLMs) to the video domain. It encodes multiple video frames into a sequence of features and feeds them into the MLLMs. The model utilizes weight fusion techniques and pooling operations to preserve learned information from image datasets during training on video datasets. Additionally, PLLaVA scales training with more data and larger language models effectively, providing detailed captions for videos.

1. Paper Link - <https://arxiv.org/pdf/2404.16994>
2. Model Link on HuggingFace - <https://huggingface.co/ermu2001/pllava-7b>

**Video LLaVA** - The architecture of Video-LLaMA consists of two branches: the Vision-Language Branch and the Audio-Language Branch. The Vision-Language Branch includes a pre-trained image encoder, a position embedding layer, a video Q-former, and a linear layer to process visual inputs. The Audio-Language Branch involves an Audio Q-Former and linear layers to handle auditory inputs.

1. Paper Link - <https://arxiv.org/pdf/2306.02858>
2. Model Link on HuggingFace - <https://huggingface.co/LanguageBind/Video-LLaVA-7B-hf>

**Current roadblock –** Tried implementation of PLLaVA on GCP instance with a T4 (Turing architecture GPU), but was faced with issue of incompatible architecture due to usage of **FlashAttention 2** which is a memory optimization library being used in all Multimodal LLMs. FlashAttention 2 requires access to at least an Ampere architecture Nvidia GPU like the A10, A100, H100, etc., or RTX 4000 series cards.

## Speech to Text

**Speech to Text (Mridul plox add your ideas here)–**

**OpenAI Whipser –**

# Datasets

1. [MBZUAI/VideoInstruct-100K](https://huggingface.co/datasets/MBZUAI/VideoInstruct-100K?row=1) - VideoInstruct100K is a high-quality video conversation dataset generated using human-assisted and semi-automatic annotation techniques.
2. [TGIF-QA](https://github.com/YunseokJANG/tgif-qa) - The TGIF-QA dataset contains 165K QA pairs for the animated GIFs.
3. [MSVRTT](https://cove.thecvf.com/datasets/839) - A large-scale video benchmark for video understanding, especially the emerging task of translating video to text. This is achieved by collecting 257 popular queries from a commercial video search engine, with 118 videos for each query. In its current version, MSR-VTT provides 10K web video clips with 41.2 hours and 200K clip-sentence pairs in total, covering the most comprehensive categories and diverse visual content, and representing the largest dataset in terms of sentence and vocabulary. Each clip is annotated with about 20 natural sentences by 1,327 AMT workers.