I read these papers over the past week to understand the transcription of amultimodal model into an LLM and seeing how vision and machine learning are being used together .

The first paper I looked at was VideoLLM : Modeling Video Sequences with Large Language modules

The experiment is as follows : VideoLLM is verified on 8 video understanding tasks across 4 datasets:

- Online Action Detection,
- Action Segmentation
- Temporal Action Detection focus on detecting and recognizing actions and their temporal boundaries.
- Online Captioning generates textual descriptions of video content.
- Highlight Detection identifies exciting parts and generates summaries.
- Action Anticipation and Long-term Anticipation predict future actions and content in advance, respectively.
- Moment Query quickly retrieves specific segments or events in a video.
- Nature Language Query localize a temporal segment through a textual question.

Key takeaways:

each short video clip is tokenized using corresponding audio and video encoders and then sequentially processed by the LLM.

- Semantic Translator: Transfers visual semantics to language semantics, aiding in the translation between modalities.
- Decoder-only LLM: Serves as a generalist video sequence reasoner for various video sequence understanding tasks, integrating the translated semantics for reasoning and understanding.

The second paper I looked at was GPT-4V(ision) System Card OpenAI

- Talks about the safety nets in GPT-4V and how the guardrails are being set to discourage dissemination of illicit information through their own product/llm.
- Discusses the ability of the model to work really well with people recognition, advanced OCR capabilities and also works really well in CAPTCHA breaking which is neat.
- Discusses its performance on different areas around which the guardrails are set to substantiate the requirement of actually putting in the guardrails since they know the models potential [See page 3 and page 4].

From the text in the paper : “”model can be used to generate plausible realistic and targeted text content. When paired with vision capabilities, image and text content can pose increased risks with disinformation since the model can create text content tailored to an image input”