Image and Video Captioning

- This project is about developing an IoT device capable of providing image/video specific and semantically correct captions

Image Captioning

- Basically the idea is to give a semantically correct description of the contents of the image using a Machine Learning Model
- More advanced versions of this model can also comment on the relative distances of different objects in the image using
- Famous datasets used for Image captioning are:
  - **COCO** dataset which is Microsoft's Common Objects in Context dataset COCO Dateset
  - **Flickr30k** dataset which contains relatively fewer images (30k images) Flickr30k Dataset
- Datasets that could be used for depth perception are
  - **DIODE** which is Dense Indoor and Outdoor Depth Dataset DIODE Dataset
- Some Popular implementations of Image Captioning are
  - **Lavis** developed by Salesforce which internally uses the **BLIP** model
  - huggingface/transformers which also uses the **DPT** model for depth estimation

Video Captioning

- The idea of video captioning extends the idea behind image captioning and allows a model to give a semantically correct description of the collection of frames constituting the video clip
- Video2Description
- Famous Datasets for video captioning are:
  - **MSVD** dataset (Microsoft Research Video Description Corpus)
  - **MSR-VTT** dataset MSR-VTT
- Some popular implementations of Video Captioning include:
  - VALOR
  - mPLUG-2

Deployment

- Since the initial idea is to develop a wearable device which can do these captioning operations, the size of the device might limit the computing power of the device and hence the model can be deployed on the Internet and which is then exposed to the device using an API.
- This device will also require a camera module to capture the images/videos and a wifi module to transfer these to the API and an amplifier module to relay the text output of the image/video captioning process.
- Further components required for this project are yet under consideration

References

- A Comprehensive Survey of Deep Learning for Image Captioning
- BLIP-2
- DPT
- Generative Image-to-text transformer
- mPLUG-2
- VALOR