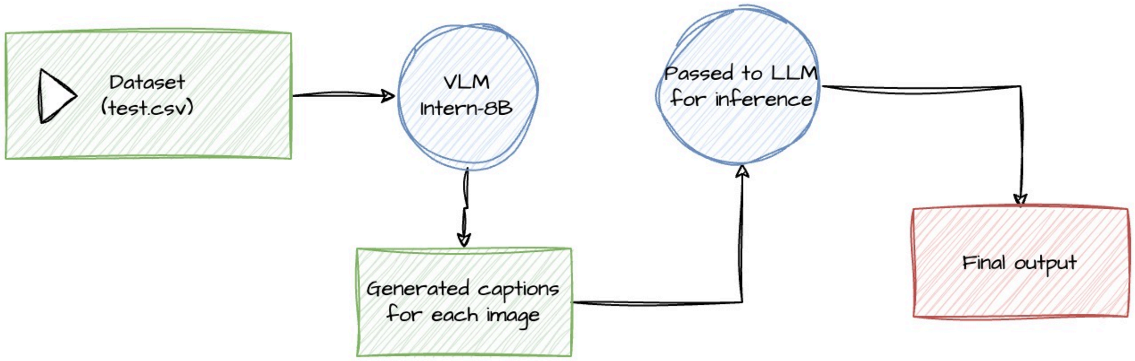
Amazon ML challenge

Team: NextTechLab

ML Approach:



Inference:

- Extract images and entity_name, values from the dataset.
- InternVL 8B 4 bit quantization: curate caption using input image and ‘entity_name’.
- Llama 70b 4 bit quantization: extracting ‘entity_value’ from the caption.
- Run inference on 131k datapoints from ‘test.csv’.
- Use Multiple GPU instances used.
- Exported the output a CSV.

GPUs used:

- Used GPUs on Cloud
- 1 NVIDIA A100 SMX
- 3 NVIDIA L40s

ML models used:

InternVL2 8B

- Vision Language Model
- 4 bit quantized for increased speed and lower memory usage.
- captioning the input images provided the ‘entity_name’.

LLama 3.1 70B

- 4 bit quantization for increased speed and lower memory usage.
- To extract ‘entity_value’ outputs from VLM captions provided the ‘entity_name’.

Experiments

1. ### Bridge Tower + decoder

   BridgeTower MLM used as an encoder to extract image and text features and train a decoder to map values to the `entity_values`.

2. ### OCR:

   train TrOCR to extract entity_values from the given image.

3. ### VLM

   Using Vision Language models to generate captions for the provided input image and entity_name.

4. ### Regex

   for extracting entity_values from VLM caption outputs.

Conclusion

Leveraged a vision language model (internvl2-8b) to generate captions and a language model (llama3.1-70b) to extract ‘entity_values’ from the said caption, the process was accelerated by the use of cloud GPUs. This provided a robust pipeline to extract ‘entity_values’, from the given data.