**Introduction**

The vision language models are getting increasingly popular in the recent times with the development in sectors like Robotics. Vision language models are multimodal models that can learn and output both text and image inputs. Their tasks include image recognition, visual question answering, image captioning, segmentation, giving text output of images, etc. These vision language models give boxes around an image or segment the object based on the requests from the user of a particular image.

With the increase in demand for VLM’s surging rapidly, the ability for these models to identify and understand image and language inputs resulted in training them on huge datasets. As a result, the models’ size is increasing which thereby is affecting the performance and analysis of data. Due to large size, the inference time of the model to output information has increased as well. Therefore, optimization of these models can improve the model’s performance efficiency, reducing latency and enabling real time performance. Moreover, by reducing the size and memory usage, these models can be used to deploy on devices with limited computational resources.

There are a variety of optimization techniques available to simplify the model while retaining its core functionality, including …

For this research, we have chosen a state-of-the-art open-source vision language model i.e. Microsoft Florence 2.