## Observations

* NNAPI doesn’t support certain layers of yolov5 and yolov7 -> only partial GPU execution possible
* Pytorch uses the NCHW input format whereas Tensorflow uses the NHWC format. Openvino allows for a seamless conversion without the need of additional transpose layers. Compare performance between two export methods?
* Conclusions: small models run faster on CPU, poor GPU interfaces for mobile devices make ML acceleration difficult task, speed-performance tradeoff is worth it when dataset is brilliant (see yolov5n vs yolov5s)
* PyTorch models are saved in FP16 by default. ONNX models export at FP32 by default. If you want an FP16 ONNX model you can export with --half flag. That’s why .onnx format is twice the size of .pt format.
* NNAPI inference time is 2x of CPU inference time (Independent of device). NNAPI is redundant since performance worse. Memory copy overhead?
* On the SDW4100 inference time scales proportionally with image input size.

## To-do

* Related Work done
* Update inference times done
* Check again if YOLOv7 works done
* Push git
* Presentation