SAVA YOLO v8 inference performance validation on Jetson

Description: This project has the objective of thoroughly testing the performance of the YOLO v8 model trained on the custom SAVA dataset. The model will be tested on the test set. The purpose of the project is to understand how the model performs under different setups, both in terms of accuracy (with respect to the experiment run on the Desktop counterpart) and in terms of inference speed.
The following setups need to be tested:

- Inference speed at different model sizes
- Inference speed and accuracy drop with different (standard) compression methods (e.g. quantization FP16 vs INT8)
- Inference speed using different model compilers: onnx, tensorRT, torch compile
- Inference speed improvement at different batch size vs single image inference
- Largest batch size that fits in memory for different model sizes and compression methods

For the testing, the model will be deployed in a Jetson AGX Orin.

Keywords: Model compression, YOLO, Inference speed

Resources:

https://docs.ultralytics.com/
https://huggingface.co/docs/optimum/en/concept_guides/quantization
https://pytorch.org/docs/stable/quantization.html
https://github.com/NVIDIA/TensorRT
https://onnxruntime.ai/

Required level:
/

Contact: fabmo@dtu.dk, ronjag@dtu.dk, chkoo@dtu.dk