Binary quantization: how much does it affect performance and accuracy?

Description: Quantization involves simplifying a neural network by reducing the number of bits used to represent its parameters, making it more efficient. This project will explore the use of binary quantization to reduce the computational and memory requirements of vision models. The focus will be on implementing binary quantization in models like ResNet or MobileNet for a task of choice like object detection, classification or segmentation. The goal is of optimizing these quantized models for deployment on resource-constrained robotic platforms, such as Raspberry Pi or Jetson Nano, without sacrificing (too much) real-time performance.

Keywords: Model quantization, Vision Models, Binary quantization

Resources:

https://huggingface.co/docs/optimum/en/concept_guides/quantization
https://huggingface.co/blog/embedding-quantization
https://weaviate.io/blog/binary-quantization

Required level:
Master student with experience in computer vision, deep learning, and Pytorch.
You should have trained deep learning model before.

Contact: fabmo@dtu.dk