Quantization in Vision Models: which architecture is most resilient?

Description: Quantization consists in reducing the precision of a model's weights and activations to lower-bit representations to save memory and computation. This project aims to investigate the effects of different quantization techniques on the performance of deep learning models in robot vision systems, focusing on a comparison of the effect among different architecture types (CNNs vs Transformer-based). The focus would be on understanding the trade-offs between model size reduction, energy efficiency, and visual perception accuracy.

Keywords: Model quantization, Vision Models, CNNs, Transformers

Resources:

https://huggingface.co/docs/optimum/en/concept_guides/quantization
https://pytorch.org/docs/stable/quantization.html
https://arxiv.org/pdf/2010.11929

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