TECHNISCHE UNIVERSITÄT
IN DER KULTURHAUPTSTADT EUROPAS
CHEMNITZ

Fakultät Informatik
Professur Neurorobotik

Master Thesis Description – Sajed Golabpars

Topic: Domain-Adaptive Quantization of Vision Transformers for Surgical Instrument Detection on Edge Devices

Description:

The aim of this work is to enable the deployment of large Vision Transformer (ViT) models for surgical instrument detection on resource-constrained edge devices through post-training quantization/Quantization-aware training. You will begin by fine-tuning a state-of-the-art ViT-based object detector (e.g., DETR, ViTDet) on the **EgoSurgery-HTS** dataset—a challenging egocentric video dataset featuring 14 surgical tool classes captured from real open surgeries. The core of the thesis is a systematic investigation of post-training quantization strategies to compress this trained model for deployment on the **NVIDIA Jetson AGX Xavier**. You will study, implement, and evaluate different quantization approaches, examining how each handles the unique challenges posed by surgical data—particularly extreme activation outliers caused by metallic instruments and the domain shift between generic pre-training and the surgical environment. The quantized models will be deployed and bench-marked across multiple bit-widths (8-bit to 3-bit), analyzing detection accuracy, inference latency, memory footprint, and energy consumption. The thesis will culminate in a comprehensive analysis of the trade-offs between compression and performance, providing practical insights for deploying surgical AI on edge hardware.

Motivation:

Deploying advanced AI for real-time surgical assistance in operating rooms is currently held back by a fundamental problem: the most accurate models for detecting surgical instruments are too large and computationally heavy for the portable edge devices that must be used due to privacy and bandwidth constraints. While Vision Transformers offer state-of-the-art accuracy, their size is prohibitive for edge hardware. Furthermore, the visual domain of surgery—with blood, tissue, and metallic glare—is vastly different from the generic datasets these models are trained on, causing performance to drop during deployment. Recent advances in post-training quantization offer a path to compress these models, but their effectiveness on the unique challenges of surgical video remains unexplored. This project aims to bridge this gap by designing a quantization method specifically tailored for the surgical domain.

Tasks:

- **Comprehensive literature review:** Conduct a literature research on Vision Transformer architectures for detection, fundamentals of post-training quantization, and existing quantization methods for ViTs.

- **Dataset Preparation & Baseline Fine-tuning:** Download and preprocess the EgoSurgery-HTS dataset for the detection task. Fine-tune a selected ViT-based detector (e.g., DETR, ViTDet) to establish a strong, full-precision baseline model and its accuracy (mAP) on the surgical data.

- **Framework Design & Core Algorithm:** Analyze the EgoSurgery-HTS dataset to identify its specific challenges (outliers, domain shift, class imbalance). Based on this analysis and literature insights, design a domain-adaptive quantization strategy. The approach must intelligently handle outliers and adapt to the surgical domain to preserve accuracy at low bit-widths

- **Evaluation & Analysis:** Implement the proposed quantization framework in **PyTorch**. Apply it to quantize the baseline detector to multiple bit-widths (8-bit, 6-bit, 4-bit, 3-bit). Test the quantized models for detection accuracy on the EgoSurgery-HTS test set. Export the quantized models and deploy them on the NVIDIA Jetson AGX Xavier using the **TensorRT** framework. Measure and record key performance metrics: inference latency (ms), throughput (FPS), memory usage, and model size. Conduct ablation studies to prove the contribution of key components in the proposed framework and perform a class-wise analysis to understand which instruments are hardest to quantize.

- **Documentation and Code:** Compile the work into a final report and presentation and upload the implementation/code on TUC cloud (link will be provided).

Further Links:

- Dataset: EgoSurgery-HTS
- Core Reference Papers:
  - RepQ-ViT: Scale Reparameterization for Post-Training Quantization of Vision Transformers
  - Domain aware post training quantization for vision transformers in deployment
- Hardware: NVIDIA Jetson AGX Xavier

Requirements:

- Fundamental knowledge of deep learning and computer vision
- Basic knowledge of model compression techniques.
- Good to have – prior practical experience with edge devices (Jetson, Raspberry Pi)