ARIA
APPLIED RESEARCH IN ACTION

Towards Intelligent Edge Computing

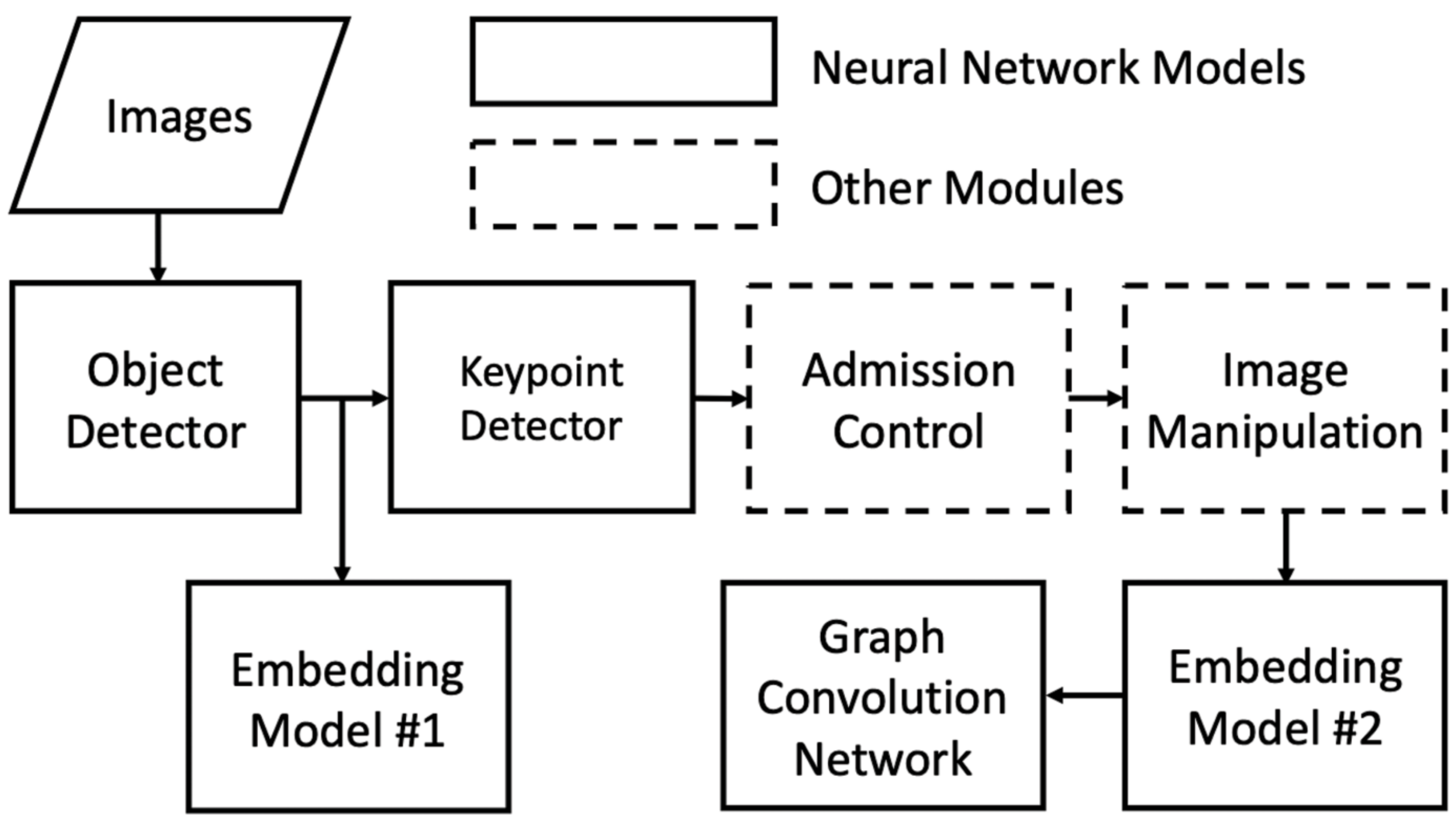
From Research Proof-of-Concept Server-based Pipeline to on-Device Deployment: A Case Study

Ke Zhao

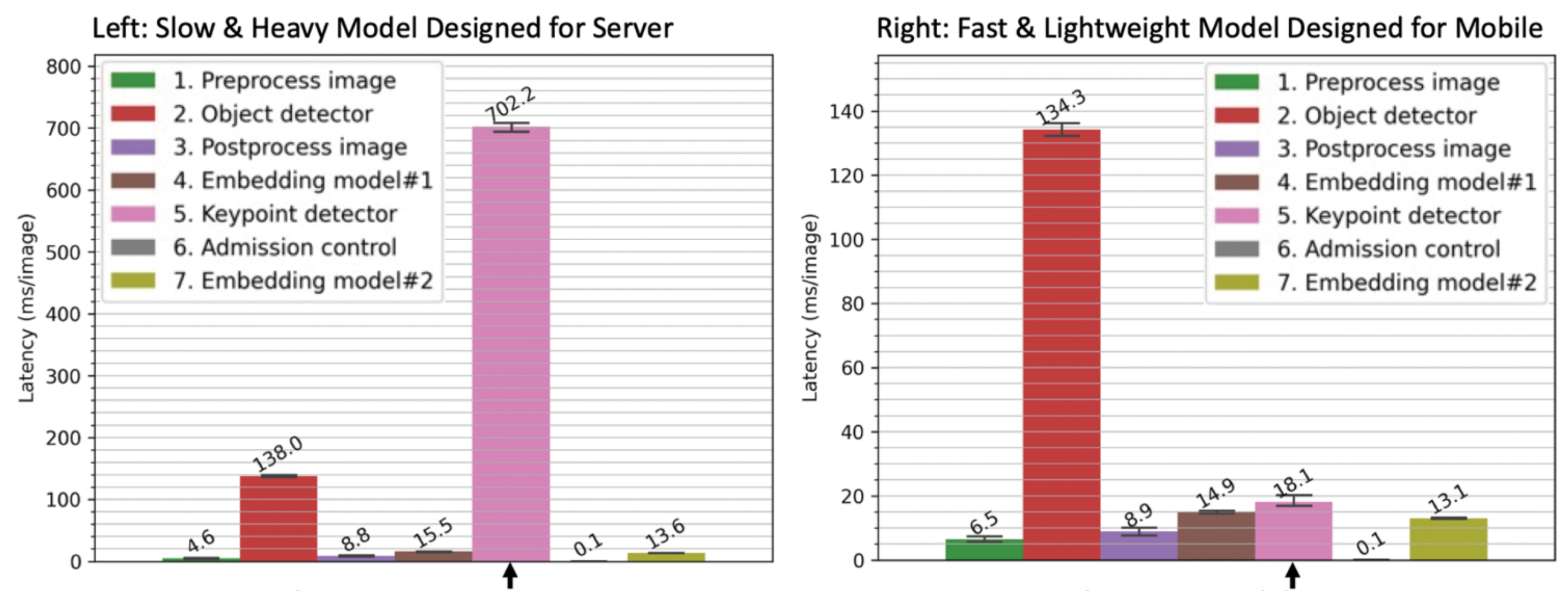
Nandita Vijaykumar
ACADEMIC SUPERVISOR

Iqbal Mohomed
INDUSTRY SUPERVISOR

Pipeline Overview



Per Image Computation Latency for Keypoint Detection Models with Two Different Architecture


Left: Slow & Heavy Model Designed for Server

Right: Fast & Lightweight Model Designed for Mobile

Architecture: KeypointRCNN
Model Format: FP32, TorchScript
Device: Samsung S23 Ultra
Latency: 702.2 ms/image

Architecture: MobileNetv3
Model Format: FP32, TFLite
Device: Samsung S23 Ultra
Latency: 18.1 ms/image

Take-away: Rearchitecting model to eliminate unnecessary computation results in far more latency savings than can be expected with model optimization.

PROJECT SUMMARY

The rapid advancement of deep learning has led to an increasing number of neural network applications in people's daily lives, raising data privacy concerns as data is often uploaded to server-run models. Mobile deployment of machine learning (ML) models on users' devices offers a promising solution as the computation only happens locally to preserve data integrity. However, deploying ML models on edge devices presents unique challenges. First, contemporary state-of-the-art models for various tasks often prove too large to execute directly on edge devices, which typically possess limited memory and computational capabilities. Secondly, mobile devices frequently feature distinct sets of hardware accelerators, such as Digital Signal Processors (DSP) and Neural Processing Units (NPU), each with its own set of assumptions regarding input data. Lastly, the variability in ML model runtime implementations and available operation sets across diverse platforms poses a non-trivial task in achieving consistent outputs. This project tackles these challenges by adapting a server-side proof-of-concept ML pipeline that contains multiple distinct-tasked models into a mobile application. Techniques such as model quantization, re-architecture, and continuous consistency checks were employed. The study demonstrates the feasibility of deploying complex ML pipelines on edge devices, shedding light on mitigating constraints posed by device limitations.

SAMSUNG

Computer Science
UNIVERSITY OF TORONTO

Master of Science in
Applied Computing