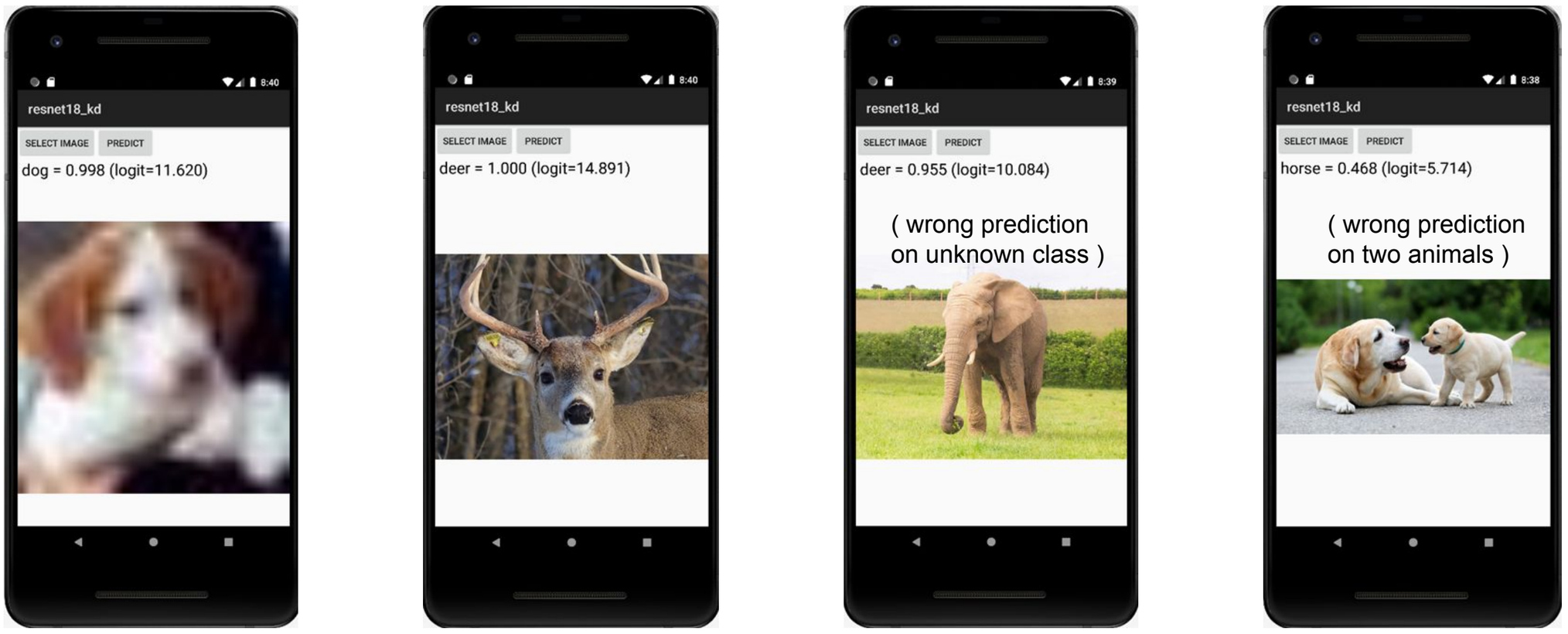
NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

School of Computer Science and Engineering
College of Engineering

Machine Learning Model Compression Techniques and Deployment on Android Platform

Student: Jin Chengkai Supervisor: Assistant Prof Zhao Jun

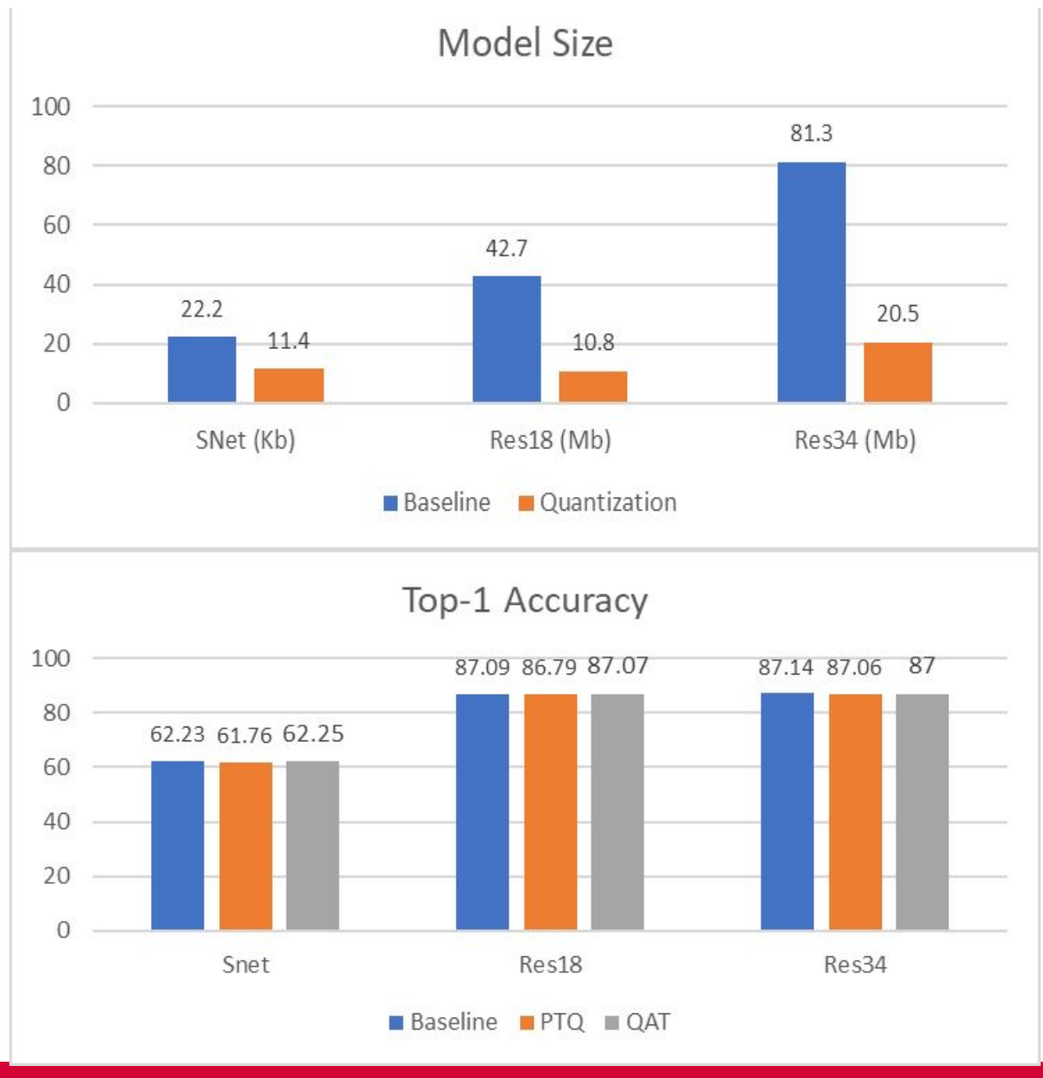


Project Objectives:

- Provide a comprehensive survey of different model compression techniques on the image classification task.
- Compare classic approaches based on quantization and knowledge distillation
- Offer a hands-on example of deploying a deep learning model on an Android application using NCNN framework.

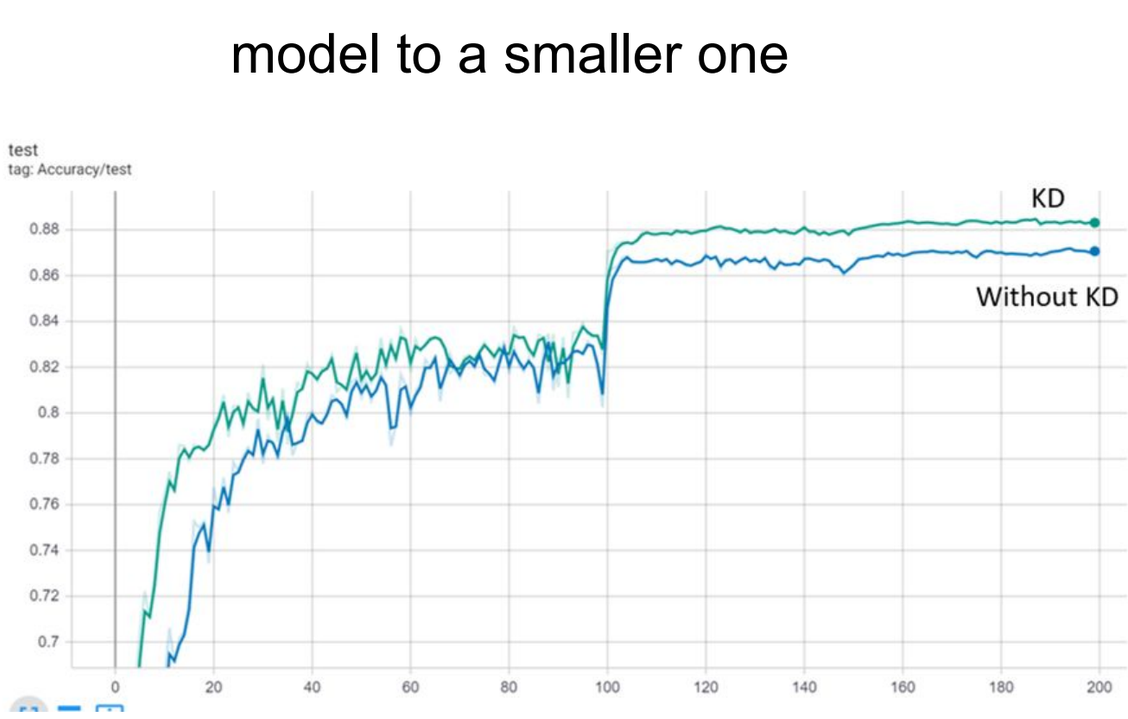
Model Quantization:

- reduce model size without significant accuracy loss



Knowledge Distillation:

- transfer knowledge from a larger model to a smaller one



distilling from ResNet34 -> ResNet18