Project proposal for Benchmarking Inference Neural Networks on mobile devices

Sobirdzhon Bobiev
Innopolis University
Innopolis, Russia
sobir.bobiev@innopolis.university

Supervisor: Adil Mehmood Khan
Innopolis University
Innopolis, Russia
a.khan@innopolis.ru

ABSTRACT

The following questions will be answered in this proposal:

1. What project will you be working on?
2. What are your goals in completing this project?
3. What is your motivation for working on this project?
4. Why should others care about your project?

1

During this work I will test one of the Machine Learning frameworks for mobile devices (TFLite, MNN, MACE) with various neural network architectures and provide a benchmarking results focused on individual ANN layer operations (Conv, Softmax, Pooling, etc.).

2

The goals of this project:

- profiling the performance of different ANN architectures with respect to their individual layers(Conv, Pool, Softmax, etc.) on CPU and GPU (OpenGL, Vulkan, OpenCL);
- collecting the performance results and perform an analysis;
- find optimal parameters e.g. kernel size, filters, etc..

3

It is a great opportunity to understand how ANN computations are performed on mobile devices and to explore where the overhead occurs. Doing this research I will

- become familiar with one of the ML frameworks and learn about ANN architectures and how to generate them automatically;
- learn how to deploy ANN to mobile devices;
- visualize the collected data and infer relationships between layer types and computation time.

4

The apps in mobile devices which use ANNs are very time-sensitive. Particularly object detection and image segmentation models may not perform in a timely manner. Therefore, we need to analyse the performance at each layer of these deep networks and get some insights on which hyper-parameters are best to use. The project is intended to provide performance guidelines for ML engineers targeting mobile devices.

PRELIMINARY PLAN

Week 5-6 (28/02)

- Decide what kind of models to use during the experiments. Download models from TFLite web site, visualize them using neutron app.
- Prepare a list of operations, that are used the most in these models, discuss them with the professor. Decide what hyper-parameters are to be experimented with.
- Decide which framework (TFLite, MNN, MACE) to choose.

Week 7-8 (13/03)

- Prepare testing environment;
- Create model generation scripts.

Week 9-10 (27/03)

- Create tool to benchmark generated models.

Week 11-12 (10/04)

- Parse logs and collect results.

Week 13-14 (24/04)

- Visualize and analyse results.

Week 15-16 (8/05)

Prepare final presentation.