CS-E4003 Special Assignment

Project Plan

An Implementation of Dynamic adaptive DNN surgery for inference acceleration on the edge

Rohit Raj
Student Number - **801636**
rohit.raj@aalto.fi

Supervisor: Prof. Mario Di Francesco
mario.di.francesco@aalto.fi
Department of Computer Science, Aalto University School of Science

1 Introduction:

The special assignment requires an implementation of the paper titled "Dynamic Adaptive DNN Surgery for Inference Acceleration on the Edge" [1]. This paper by C. Hu *et al.* describes a partition scheme to efficiently maximize the performance of Deep Neural Network (DNN) inference on the edge and cloud machines. The partition scheme adapts itself under light and heavy network conditions and distributes the DNN layers accordingly. More specifically, my aim in this special assignment would be to implement the Edge-Cloud DNN Inference Model - Light (ECDI-L) proposed by the authors.

2 Approach:

The approach to the special assignment and the implementation would first start with a thorough reading of the paper. The first step towards implementation would be selection of a suitable model and a Deep-Learning Framework. After this, I will start the implementation of ECDI-L scheme for DNN layers partitioning in Python 3.6.

3 Expected Outcome:

The outcome of this special assignment will be commented source code of ECDI-L scheme.

4 Credits and Grading:

The special assignment will be worth 5 ECTS and will be graded as either **PASS** or **FAIL**.

5 Timeline:

The expected timeline is as follows:

- (24/01/2019)Complete the reading of the research paper.
- (31/01/2019) Finalize the choice of DNN Model and Framework
- (22/05/2019) Complete the implementation of ECDI-L in python
- (31/05/2019) Submission of source code and required documentation to the supervisor.

Additionally, I will report the progress of the implementation to the supervisor weekly/biweekly personally or through email.

6 References:

[1] C. Hu, W. Bao, D. Wang, and F. Liu, "Dynamic adaptive DNN surgery for inference acceleration on the edge," in IEEE INFOCOM 2019 - IEEE Conference on Computer Communications, April 2019, pp. 1423–1431.