**Scope of work:**

Parser for deep neural network models written in PyTorch or TensorFlow

**Motivation:**

We are planning to build an inference ASIC. In order to estimate the performance, we need to evaluate our performance for various deep neural network models

**References:**

Given a reference model (for example <https://github.com/mlcommons/inference>), parse the model and extract the layers into a CSV file

**Expected delivery of this project:**

A python code that performs parsing to deep neural network models which are written in PyTorch or in TensorFlow. This parser creates a CSV file that describes the operators of each layer. Specifically:

* The type of operation of the layer (i.e. conv2, batchnorm, activation (and what type), etc.)
* The size of the layer (channel in, channel out, filter width, filter height)
* The size of the image (image in [W, H], image out [W, H])
* Operators (stride, pad, max pool, average pool, global pool, etc.)
* Size of data [bytes] generated by the layer
* And any other information related to the layer

With the CSV file we will be able to copy-paste into an EXCEL and estimate the effect and execution time in our platform. We will make the EXCEL file, it is not in the scope of this project.

**Qualification of project completion:**

To qualify, we would ask that the code written is able to correctly parse the following models: ResNet50, Yolo V3, BERT 768 (with 12 multi head attentions), RetinaNet