CS 6886: Systems Engineering for Deep Learning

Assignment 1

Inference on Inference Times

Instructions

1. This assignment is released on 4th Feb and is due on **13th Feb** (midnight)
2. The total marks (in overall weightage for course) is **10 marks** for this assignment
3. This assignment consists of two parts.
    a. Part A is compulsory and is worth 10 marks.
    b. Part B is optional and is worth 2 bonus marks
4. You are required to submit a **PDF report** with contents as described in each of the questions.
5. The report is to be submitted on Google Classroom.
6. Copying of code or parts of the report is disallowed.
7. You are allowed to discuss with each other and refer to sources online. You are required to clearly state any such collaboration or references in the report.

Part A (10 points)

Step A1 (2 points)

1. Go through the MLPerf benchmark paper here.
2. In the report, list all the 5 models which form part of the MLPerf Inference benchmark set given here.
3. Obtain trained PyTorch networks for these 5 models. In the report, cite the source for each of them.
4. Test each of the 5 networks for a sample input (from the dataset listed in the MLPerf page) and demonstrate that the model works as expected by including sample output in the report.

Step A2 (4 points)

5. Export the PyTorch networks to ONNX networks based on the instructions described here.
6. Run the ONNX models using the ONNX runtime on your computer and demonstrate again that the models work as expected by including sample output in the report.
7. Go through this repository on running ONNX in a web browser through ONNX.js.

8. Run each of the models on a web browser on your computer and include screenshots of the console window in the report to demonstrate that the models are working correctly.
9. Use `console.time` and `console.timeEnd` calls in Javascript to compute the time taken for inference.
10. Figure a way to run the model inference on the same input 30 times and report mean and variance for each model on your computer.

Step A3 (4 points)

11. Run each of the models on 5 different devices spanning laptops, mobile phones, and/or servers. List all devices you tried and those that worked.
12. List down the main specs of the 5 devices (CPU cores, core frequency, main memory size, cache size, presence of accelerator).
13. For each of the models and each of the devices report mean and variance of inference time.
14. Plot the inference times as a grouped bar plot where each group is a device and different bars in a group are the inference times (mean) on each of the models.
15. Compute the Pearson's correlation coefficient between the inference times of every pair of devices. Which pairs of devices are most correlated?
16. Compute the Pearson's correlation coefficient between the inference times of every pair of models. Which pairs of models are most correlated?

Optional Part B (2 bonus points)

17. In the above analysis, the inference time corresponds to the working of the entire model. However, we would like to analyse the inference time at a higher granularity. Your goal is to understand and modify ONNX.js to report inference time for each of the individual layers. Report the modification made.
18. (Not to be reported) Once this is done, you can have fun in analysing the inference times across layers and devices.

// This is a fun assignment. Don't be like this.