The University of Texas at Austin
Chandra Department of Electrical and Computer Engineering
Cockrell School of Engineering

Spring 2023

ECE 361E: Machine Learning and Data Analytics for Edge AI
HW3 Assigned: Feb 14, DUE: Feb 23 (11:59:59pm CST)

Work in groups of two students. At the end of the PDF file, insert a paragraph where you describe each member's contribution and two valuable things you learned from this homework. Only one submission per group is required.

Introduction

In this homework, you will deploy popular deep learning models on different computational platforms (i.e., Odroid MC1 and RaspberryPi 3B+) using Open Neural Network Exchange (ONNX), one of the most used frameworks for deployment. To compare inference latency, accuracy and energy consumption during inference, you will use the CIFAR10 dataset. By working on this assignment, you will be able to:

- Understand the process of deploying PyTorch models on edge devices using ONNX (e.g., converting a model to ONNX, deploying it on real edge devices and doing inference);
- Measure latency, accuracy, energy consumption and CPU temperature variation during inference in order to understand the impact of model inference on edge devices;
- Understand the importance of designing models optimized for edge devices.

Problem 1 [35p]: PyTorch Evaluation of VGG models

Question 1: [10p] Starting with the code of VGG11 given to you, create the new VGG16 version; place it in the ***models*** folder. Use the naming convention of adding the suffix ***_pt*** to the model name (e.g., ***vgg11_pt.py*** and ***vgg16_pt.py***).

Question 2: [15p] Train the VGG11 and VGG16 models in parallel on Maverick2 using ***main_pt.py***. Follow the ***TODO*** parts from the ***main_pt.py*** code to make it work properly and then add the necessary code needed to complete ***Table 1***.

Table 1

Model	Training accuracy [%]	Test accuracy [%]	Total time for training [s]	Number of trainable parameters	FLOPs	GPU memory during training [MB]
VGG11						
VGG16						

Question 3: [10p] Draw a *single* plot that shows the test accuracy of VGG11 and VGG16 vs. the number of training epochs. Based on this plot and the results in **Table 1**, compare VGG11 and VGG16; explain which model you would choose for *training*.

Problem 2 [45p]: Deployment on Edge Devices Using ONNX

Question 1: [5p] Create a new file named ***convert_onnx.py*** that contains a function to convert the PyTorch models in the ONNX format and save the converted models using the ***_pt*** suffix and the ***.onnx*** file extension.

Question 2: [20p] Use the ***deploy_onnx.py*** file from the ***HW3_files*** folder to deploy the VGG ONNX models on the RaspberryPi 3B+ and Odroid MC1 devices. Complete the ***TODO*** parts in the ***deploy_onnx.py*** file. Add some extra code to calculate and report the accuracy of the four ONNX models on the **test dataset**. Perform inference on the *entire* test dataset available in the ***HW3_files/test_deployment*** folder on both Odroid and RaspberryPi devices; then, complete ***Table 2***.

Table 2

	Total inference time [s]		RAM memory [MB]		Accuracy [%]	
	MC1	RaspberryPi	MC1	RaspberryPi	MC1	RaspberryPi
VGG11						
VGG16						

Question 3: [20p] For the VGG11 model, draw a *single* plot of the variation of power consumption over time [s] for both devices (i.e., one curve for RaspberryPi and one curve for MC1). Then, on a different plot the variation of average temperature measurements of the CPU for each device over time [s] (i.e., one curve for RaspberryPi and one curve for MC1).

Complete ***Table 3***. Based on the plots drawn and the data collected in ***Tables 2*** and ***3*** compare the performance of the two edge devices. Which device would you prefer to use for inference? Explain.

Table 3

Model	MC1 total energy consumption [J]	RaspberryPi total energy consumption [J]
VGG11		
VGG16		

Problem 3 [20p + 10Bp]: MobileNet-v1 on Edge Devices

Question 1: [5p] Train MobileNet-v1 on the CIFAR10 dataset using the ***main_pt.py*** file and the ***mobilenet_pt.py*** model on Maverick2. Extend ***Table 1*** in Problem 1 by inserting a new row below the table with the results obtained for MobileNet-v1.

Question 2: [15p] Use ***convert_onnx.py*** (from **Problem 2, Question 1**) to convert the MobileNet-v1 model in ONNX and deploy it using the ***deploy_onnx.py*** file on both Odroid MC1 and RaspberryPi 3B+. Extend ***Table 2*** and ***Table 3*** to include (in new rows) the results you got for MobileNet-v1.

BONUS Question 3: [10Bp] Compare the data in the extended ***Table 1***, ***Table 2*** and ***Table 3*** from **Problem 2, Question 3** and **Problem 3, Question 2**. Which model delivers better results (i.e., latency, energy, accuracy) on each edge device? Explain.

Submission Instructions

Include your solutions into a single zip file named <**Group#**>**.zip**. The zip file should contain:

1. A single PDF file containing all your results, tables, plots and discussions.
2. A *readme.txt* file explaining all your items in the zip file.
3. Your code files, named suggestively (e.g., **p1_q1.py** for **Problem 1 Question 1** code).

Good luck!