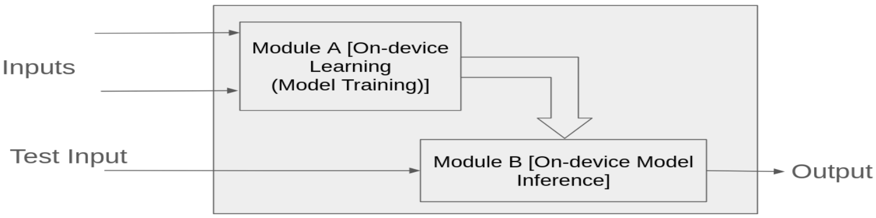
IIT JODHPUR

Minor Examination: EEL71020 Hardware Design for AI (OPEN-BOOK)

Guidelines (Total time: 120 minutes, Maximum Marks: 25):

- Please read the question paper very carefully. **NO clarification is required in any question.** In case of any doubt, assume whatever you wish to and state that in your answer.
- In case you belong to MTech(RMS)/MTech(AI)/MTech(AR/VR) batches, you can represent your design through an appropriate/equivalent C++/C description instead of describing the requested design through hardware structures/blocks. You are NOT allowed to use any library while describing the design.

1. Suppose you are employed in a bangalore-based deep tech startup, ReImagineAI. This startup comes up with an ambitious plan of implementing on-device (on-chip) model training and inference as per the high-level illustration shown below.

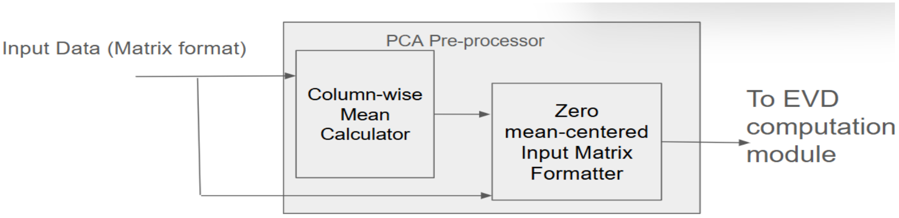

Inputs
Module A [On-device Learning (Model Training)]
Test Input
Module B [On-device Model Inference]
Output

a) Module-A involves iterative comparing inputs with some sort of threshold and then builds a learning model by utilizing mean-square error value as the loss function. Implement this loss function in hardware with proper assumptions. [3 Marks]

b) It is known that the learning model developed (with the training procedure) is a simple linear regression function ($f(x)$). If it is told to you that $f(x)$ has 14 learnable coefficients and 1 constant term. Identify the critical parameters in implementation of Module-B and then design it? Assume that Module-B has some amount of on-chip memory available inside it. Consider inputs are quantized into 16-bit short integers, take the coefficients & constant also as 16-bit short integers? [3 Marks]

2. You decide to utilize on-chip version of SVM algorithm for sensor data classification in an IoT-based system design. In order to cut down the system latency, you decide to use a mix of linear and polynomial kernels for input data transformation. Given that the relationship between input dimensions (x_1,x_2,x_3) and the transformed dimensions (z_1,z_2,z_3&z_4) are given by this function **f(x_1,x_2,x_3) = {5*x_1, x_1*x_2, x_2+3*x_3, 8*x_4-x_1+x_3*x_3}**. How would you carry out this transformation in an on-chip manner if you are instructed to minimize the resource usage in your design? [5 marks]

3. You are working in a company named as *Mewar Electronic Design* incubated at IIT Jodhpur. This company has an ongoing project to implement the popular PCA algorithm in hardware. You decide to implement the complete algorithm as per the below plan (where you first implement the pre-processing part, referred to as "PCA pre-processor" as a separate module and the eigen value decomposition (EVD) as a separate module):


Input Data (Matrix format)
PCA Pre-processor
Column-wise Mean Calculator
Zero mean-centered Input Matrix Formatter
To EVD computation module

If the input dimensions are 256 x 1024 (FP numbers), please develop the above PCA pre-processor design (block shown in the above figure) and obtain the latency values using the parameter values as listed below? What is the total resource consumption (in terms of number of functional units such as adders/multipliers etc.) of your design? [2+2+2 Marks]

Parameter	Value
Time taken for each integer multiplication	2 cycles
Time taken for each integer addition/subtraction	1 cycle
Time taken for each integer comparison	1 cycle
Time taken for each memory access from secondary memory	8 cycles
Time taken for each integer division	4 cycles
Time taken for each FP32 multiplication	4 cycles
Time taken for each FP32 addition/subtraction	2 cycles
Time taken for each FP32 comparison	2 cycles
Time taken for each FP32 division	6 cycles
Time taken for each memory access from on-chip memory	3 cycles

4. Consider that an alumni of IITJ funds you to develop custom hardware for his/her ML-based startup. As the only application of this company is image classification, you decide to utilize K-means-based clustering using Hamming distance metric. In order to calculate the Hamming distance between two strings, we perform their XOR operation, $(a \oplus b)$, and then count the total number of 1s in the resultant string. Suppose there are two numbers 11011001 and 10011101. 11011001 $\oplus$ 10011101 = 01000100. Since, this contains two 1s, the Hamming distance, d(11011001, 10011101) = 2. Design the hardware for realizing the K-means implementation if it is known that there are 5 output classes and the inputs are 32-bit integers such that the overall system latency is minimized? [4 marks]

5. You are hired as a consultant to *BharatBullsTraders*, a Noida-based high frequency trading firm. This firm engages in selling and buying stocks of various companies registered on Bombay Stock Exchange (BSE) through machine learning model building and inference. The dataset on which the company works is shown as below:

Date	Open	High	Close	Low	Adjusted Close	Volume Traded
..	..	..	..	..	..	..

The meaning of the above terms are explained as below:

Parameter	**Meaning**
Date	Date of trading
Open	Opening value of the stock on that day
High	Highest value of the stock on that day
Close	Closing value of the stock on that day
Low	Lowest value of the stock on that day
Adjusted Close	Closing price that has been modified to account for corporate actions
Volume Traded	Number of stocks traded (bought/sold) that day

As a consultant, you advise to the *BharatBullsTraders* firm that one way in which they can outshine their competitors is through implementation of the machine learning on specialized FPGAs instead of using off-the-shelf CPUs. The engineers of this firm have observed that certain ML models are providing the same performance in terms of accuracy, however, they are confused which model they should implement on the hardware. [3 + 1 Marks]
(a) What would be your advise on the choice of ML model given that FPGAs offer parallelization implementation choices but offer finite resources (LUTs/FFs/DSP slices etc.)? Why?
(b) Are all features (of the dataset as shown above) important for the analysis?