IIT JODHPUR

Minor-2 Examination: EEL71020 Hardware Design for AI (Mar'24: OPEN-BOOK)

Guidelines (Total time: 60 minutes, Maximum Marks: 15):

- Please read the question paper very carefully (both sides of this paper).
- **NO clarification is required in any question. If you ask the invigilator any question, -5 penalty would be awarded.** In case of any doubt, assume whatever you wish to and state that in your answer.
- Use of internet is NOT allowed.
- Do not write unnecessary things in your answer sheet. This may attract penalty.
- For all questions (except Question-4), please use the following parameters:

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	6 cycles
Time taken for each integer division	3 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

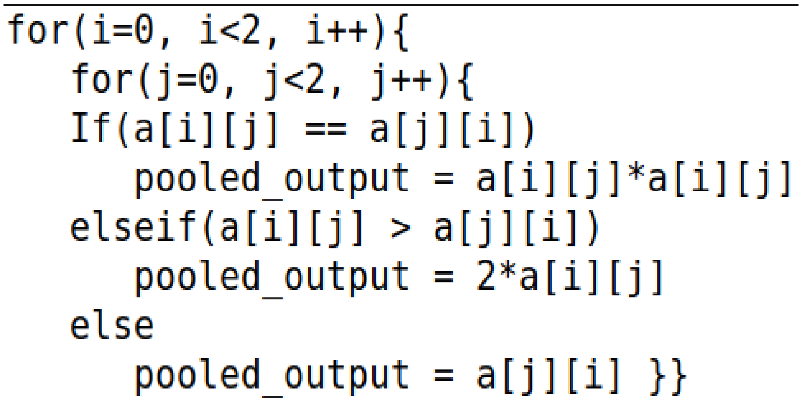
---

1. Consider that a new learning model (LM_{IITJ}) has been developed. This new model involves computing the average of 10 individual output functions (Y_j). The individual function, Y_j is given as m_jX + C_j where X represents the input. The values (m_j,C_j) are to be fetched from on-chip memory. Find out an expression for the total time (T_P) taken to compute any prediction from LM_{IITJ} given that the values (m_j,C_j) are FP32 numbers? Using the parameter values above, find the numerical value of T_P? [2 + 1]

2. Assume that the students of EEL71020 are attempting to analyze the effectiveness of 2 models for an image classification task based on neural networks. The two models (M_A and M_B) are defined as below: M_A is having 1 input layer, 1 output layer and 1 hidden layer with sigmoid activation function (after the hidden layer), whereas M_B is having 1 input layer, 1 output layer and 2 hidden layers with RELU activation function (after each hidden layer). Given that each model has fully-connected (FC) layers and all the weights are integers only, find out which one out of M_A or M_B models would provide close to real-time inference performance given that both of these models guarantee the same classification accuracy? You are provided with the information that all the respective weights are to be fetched off from the secondary memory. You can approximate the sigmoid function to second-order term. [3]

3. One of the important steps in implementation of PCA algorithm for dimensionality reduction is the calculation of zero mean centering vector. In a particular design of the computation unit for zero mean centering vector, *OutValid* signal becomes high when the result has been computed and ready to be sent to the next block of PCA algorithm implementation unit. Given that the input data is of dimension 10x800 that is present in on-chip memory, find the clock cycle in which *OutValid* is asserted (i.e., becomes 1) in these two cases: a) if it is known that all input data numbers are integers? b) if it is known that all input data numbers are FP32? [3 + 1]

4. If a very capable AI engineer from CSE, IITJ suggests to you that 2x2 pooling can be done in the following manner for implementing a CNN-based deep learning model, find the equivalent hardware implementation so that this computation can be done in a very quick manner? Is there any way to enable this pooling operation in a single clock cycle? [2.5 + 1.5]



```
for(i=0, i<2, i++){
   for(j=0, j<2, j++){
   If(a[i][j] == a[j][i])
      pooled_output = a[i][j]*a[i][j]
   elseif(a[i][j] > a[j][i])
      pooled_output = 2*a[i][j]
   else
      pooled_output = a[j][i] }}
```

5. Explain any one scenario from day-to-day life where you have felt the necessity of hardware design for AI/ML applications? [1]