IIT JODHPUR

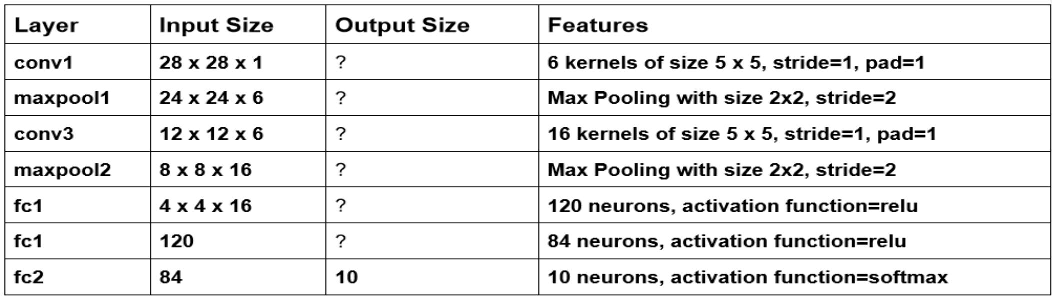
Major Examination (OPEN-BOOK): EEL71020 Hardware Design for AI

Guidelines (Total time: 120 minutes, Maximum Marks: 30):

- Please read the question paper very carefully (both sides of this paper). **NO clarification is required in any question.** In case of any doubt, assume whatever you wish to and clearly state that in your answer.
- This is an OPEN-BOOK examination. However, usage of internet/any AI tool is **NOT** allowed.
- Please write your answer to-the-point ONLY.

---

1. LeNet-5 is a popular neural network model for handwritten digits classification. The below figure presents its architecture. Find out the total number of parameters and multiply-accumulate (MAC) operations in this neural network (NN) model? What is the value of the respective performance metric if this NN model is executed on a CPU running at 2 GHz? (Hint: Try filling the values marked as ”?” first.) [8 marks = 4 + 2 + 2]



Layer	Input Size	Output Size	Features
conv1	28 x 28 x 1	?	6 kernels of size 5 x 5, stride=1, pad=1
maxpool1	24 x 24 x 6	?	Max Pooling with size 2x2, stride=2
conv3	12 x 12 x 6	?	16 kernels of size 5 x 5, stride=1, pad=1
maxpool2	8 x 8 x 16	?	Max Pooling with size 2x2, stride=2
fc1	4 x 4 x 16	?	120 neurons, activation function=relu
fc1	120	?	84 neurons, activation function=relu
fc2	84	10	10 neurons, activation function=softmax

Figure 1: LeNet-5 architecture

2. Below figure represents a scheme for efficient neural network model inference execution. Design the quantization hardware where it is known that the original FP32 weights are fetched from external DRAM and the target format is INT8? How may clock cycles would be spent by this hardware for the quantization of any one parameter? Calculate the performance speed-up when we decide to utilize INT4 quantized values? [6 marks = 3 + 1 + 2]

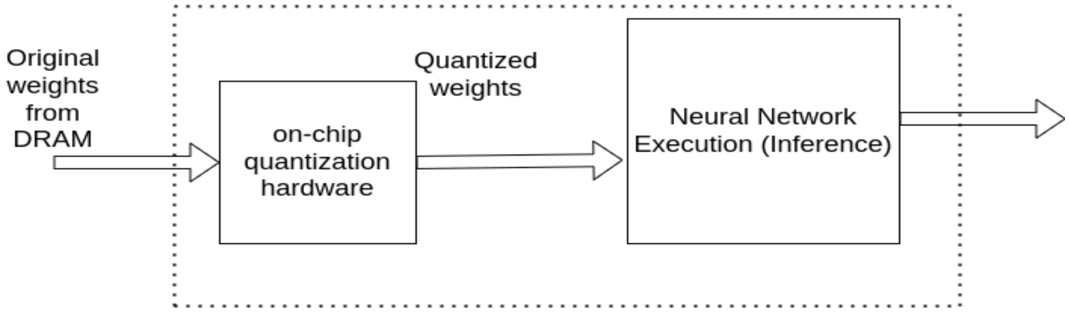

Original weights from DRAM
on-chip quantization hardware
Quantized weights
Neural Network Execution (Inference)

Figure 2: Scheme for NN implementation

3. Design a simple on-chip neural network pruning circuitry that can transform a fully-connected (FC) layer into a sparsely connected layer. Assume that pruning is based on comparison of weight values with a threshold? Please take any other assumptions that you may need. [3 Marks]

4. It is observed that a convolution layer is the bottleneck in a particular DL model execution on a CPU. We have the option of executing the computations corresponding to this layer on a specialized accelerator. Design the dataflow of this accelerator and suggest an architecture of connecting the processing elements (PE) and the input/output buffers if it is known that the image size is 7x7 and the kernel size is 3x3. You can assume the stride to be equal to 1. What happens to the dataflow in your accelerator architecture when we consider stride size to be 2? [7 Marks = 5 + 2]

5. Suppose we wish to implement the famous ML model of K-nearest neighbor (KNN) algorithm as a dedicated hardware design. The steps of this learning model are listed as below:

   - Assign a value to K.
   - Calculate the distance between the new data entry and all other existing data entries. Arrange them in ascending order.
   - Find the K nearest neighbors to the new entry based on the calculated distances.
   - Assign the new data entry to the majority class in the nearest neighbors.
   - Keep repeating the above till all data entries are finished.

   With proper assumptions about the above model, please design an on-chip optimized custom digital logic implementation of this model? [4.5 Marks]

6. What did you learn from this course? [1.5 Marks]