9

Hardware Architecture for Deep Learning - CS6490. Spring 2024-25.

Dept. of CSE, IIT Hyderabad

Quiz-1, Set-B **Time: 10 minutes**

Total marks: 10

Name: Pandranga Adithya Sriram

Roll number: EE22BTECH11039

---

1. The depth (z-dimension) of a kernel in a CONV layer is same as (choose correct ones): [1]
   - a. ✓ The number of input channels
   - b. ~~The number of output channels~~
   - c. The number of columns of the output
   - d. The number of rows of the input

2. The number of kernels/filters in a CONV layer of a CNN is same as (choose correct ones): [1]
   - a. The number of input channels
   - b. ✓ The number of output channels
   - c. The number of columns of the output
   - d. The number of rows of the input

3. Consider a CONV layer of a CNN with following characteristic: Input: 192x192x7, #filters=48, filter size = 3x3, stride=1. Find the dimensions of the output. [2]

Stride = 1 ⇒ Output size (length/width) = 192 − 3 + 1 = 190

Depth = ~~7x48~~ channels Output channels = 7 per filter

48 / 1 / 56 / 28 / 336

∴ with 48 filters, dimensions are 190 × 190 × 336

4. Consider an FC layer of a CNN with following characteristics: Input: 7x7x96, #filters=2048. Find the dimensions of the filter and the output. [2]

Filters : 7x7 each (since it is FC) and 96 channels

Output Dim : 7 − 7 + 1 = 1 × 1

↳ 1 × 1 × 2048

2

5. Perform 2-D convolution of the given 4x4 input with the 2x2 kernel. Assume stride=2. [4]

Input:

5	7	2	1
6	1	2	5
1	2	0	2
11	3	2	9

Kernel:

-1	2
-3	4

Output size $= \left(\frac{4}{2} + \frac{2}{2} - 1\right)$
(length or width)
$= 2$

$\therefore$ 2x2 matrix

$out[0][0] = 5(-1) + 7(2) + 6(-3) + 1(4)$
$= -5 + 14 - 18 + 4$
$= -5$

$out[0][1] = 2(-1) + 1(2) + 2(-3) + 5(4)$
$= -2 + 2 - 6 + 20$
$= 14$

$out[1][0] = 1(-1) + 2(2) + 11(-3) + 3(4)$
$= -1 + 4 - 33 + 12$
$= -18$

$out[1][1] = 0 + 2(2) + (-3)(2) + (4)(9)$
$= 4 - 6 + 36$
$= 34$

4

$\therefore$ output $= \begin{bmatrix} -5 & 14 \\ -18 & 34 \end{bmatrix}$

(9)

Hardware Architecture for Deep Learning - CS6490. Spring 2024-25.

Dept. of CSE, IIT Hyderabad

Quiz-2, Set-A

Total marks: 10 **Time: 10 minutes**

Name: Pandiangi Aditya Sriram

Roll number: EE22BTECH11039

Choose all correct choices for the MCQ.

1. A systolic array enables: [2]
   - a. Parallel processing
   - b. Data reuse
   - c. Run-length compression
   - d. Simpler design

2. Which dataflow is depicted by the given code snippet for a 1-D convolution: [1]

```
for (i=0; i<8; i++){
    out[i] = 0;
    for (j=0; j<3; j++)
        out[i] = out[i] + in[i+j]*w[j];
}
```

out, in, wt represent the output, input, and weight matrices, respectively.

   - a. Output stationary
   - b. Input stationary
   - c. Weight stationary
   - d. Row stationary

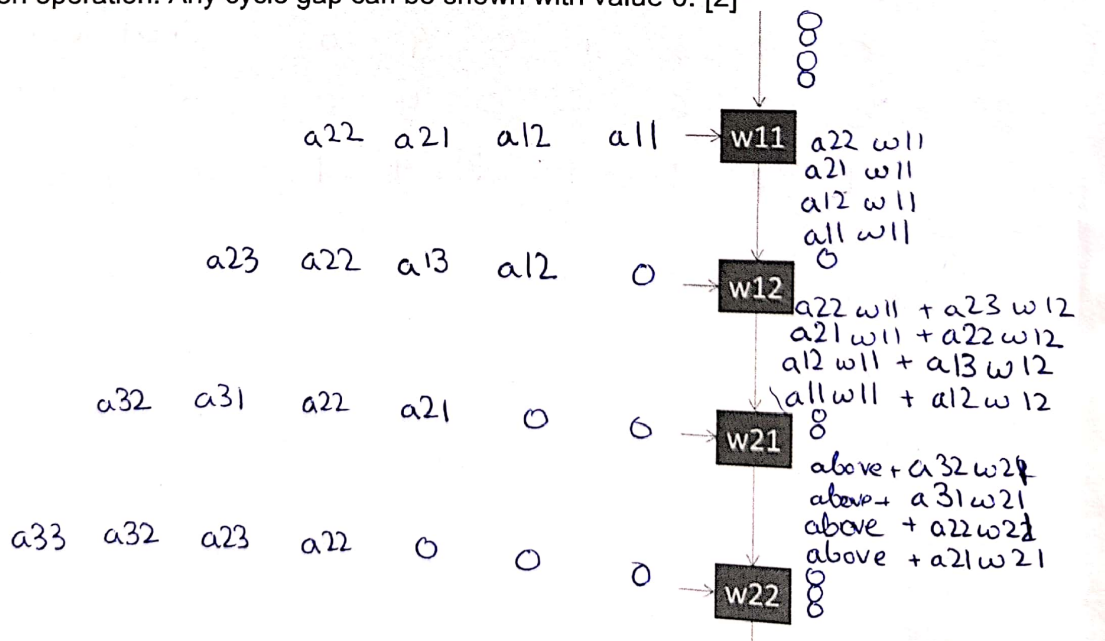
3. Given the systolic design as shown, mark the values to be fed to horizontal and vertical inputs to perform the convolution operation. Any cycle gap can be shown with value 0. [2]

Input:

a11	a12	a13
a21	a22	a23
a31	a32	a33

Kernel:

w11	w12
w21	w22



2

a22 w11 + a23 w12 + a32 w21 + a33 w22

⋮

4. An application takes 15 seconds time to execute on a multi-core machine but takes 150 seconds on a single-core machine. What could be the maximum value of the time taken by the sequential (non-parallelizable) portion of the code on the single-core machine? [2]

By Amdahl's law, (let f be sequential portion)

$$\frac{150}{15} = \frac{f + 1-f}{f + \frac{1-f}{n}} \quad \text{as } n \to \infty \text{ (infinite cores)}$$

$$\Rightarrow 10 = \frac{1}{f} \Rightarrow f = 0.1$$

$$= 10\% \equiv 15 \text{ seconds}$$

5. In the Eyeriss accelerator discussed in the class, data gating helps to improve: [1]
   a. Energy consumption
   b. Execution time or speed
   c. Area

6. Compress the given data as per the run-length compression scheme of Eyeriss. Show the value of term-bit as well. [2]
   12, 0, 0, 0, 1, 9, 8, 0, 0, 0, 0, 0, 0, 112, 4, 0, 2, 0, 7, 0, 0, 0, 9.

   The encoding format is as shown below:

Run Level Run Level Run Level Term

5b	16b	5b	16b	5b	16b	1b	
0	12	3	1	0	9	0	(next row)
0	8	6	112	0	4	0	(,,)
1	2	1	7	3	9	1	

Hardware Architecture for Deep Learning - CS6490. Spring 2024-25.

Dept. of CSE, IIT Hyderabad

Quiz-4 **Time: 12 minutes**

(8)

Total marks: 10

Name: Pandranki Aditya Sriram

Roll number: EE22BTECH11039

---

1. Which of the following is/are true for the Channel Gating network (CGNet)? Give 1-2 line reasoning for your choices. [2]
   a. It deploys a static pruning scheme
   b. It provides a fixed speedup, independent of the input image
   c. It deploys a dynamic pruning scheme
   d. The speedup depends upon the provided input image

a. False. The pruning is dynamic based on ~~input image an~~ PSUM's generated till a certain fraction of input channels. ~~while~~ during inference.

b. False. Speed-up is variable and depends on ~~to~~ the fraction of discriminatory information in input (lower fraction → higher speedup).

c. True

d. True.

2

2. DNNExplorer uses a custom pipelined structure for initial few layers, but a generic structure for later layers. Why? [3]

→ The computation-to-communication ratio (CTC) has very high variance for earlier layers in a CNN (in general), across several orders of magnitude. This is true even for a fixed set of parameters with variable input.

→ However, for deeper layers, this variance decreases.

→ ~~Thus~~ Having a general structure throughout makes it ~~it~~ so that there is a potentially huge wastage of resources of accelerator (asr) higher latency (in case of high communication), while specific structure does not allow for usage for various kinds of networks (generality).

3

A balance is struck by using a pipelined and generic structure each having roughly equal latency.

low variance CTC thus can be generalized well

max CTC / min CTC ratio is small

3. We observed that a few of the compressed networks are able to provide higher accuracy than original uncompressed networks. What makes this possible? [2]

In networks such as mobile net, accuracy remains similar even after a large amount of compression

- Overfitting may reduce for compressed networks with less parameters.
~~Sometimes, the depth-wise~~
For eg seperating depth-wise and point-wise convolution leads to a similar structure of different kernel channels, which may resolve input features better, as less parameters are easier to train.

~~Training also~~ t

4. Explain the significance of parameters S_{1x1}, e_{3x3}, and e_{1x1} in SqueezeNet. Is there any relationship that must hold between these parameters for an effective compression? [3]

S_{1x1} indicates number of 1x1 squeeze kernels in a fire module.

e_{3x3} indicates " " 3x3 expand " " "

e_{1x1} " " " 1x1 " " " "

- To achieve effective accuracy, the model must compress data more aggressive a later point of the network, so that the network can gather most of the useful information ~~early~~ by then.

2 ✓

Thus, the fraction of squeeze kernels must increase for deeper layers, ~~and expand ker~~ and should be low for earlier layers.

~~3x3 kernels must also be~~

6.5

Hardware Architecture for Deep Learning - CS6490. Spring 2024-25.

Dept. of CSE, IIT Hyderabad

Quiz-5

Total marks: 10 **Time: 12 minutes**

Name: Pandrangi Aditya Sriram

Roll number: EE22BTECH11039

---

1. Which of these choices represent a tiny device considered for the MCUNet work? [1]
   a. SRAM: 1GB, Flash: 4GB
   b. SRAM: 256KB, Flash: 1MB
   c. SRAM: 320KB, Flash: 100MB
   d. SRAM: 256KB, Flash: 2MB

0.5

2. Assume a hypothetical tiny device with infinite sized Flash (regular sized SRAM). In what aspects can such a device make inferencing better than a real tiny device? What limitations of existing tiny devices cannot be resolved by such a hypothetical device. [4]

Flash is a read-only memory for practical purposes on accelerators.

Advantages of inf flash :

① ~~All the~~ The complete code can be converted to instructions & stored. All loops can be unrolled. This significantly reduces loop overheads (branch)

② ~~Infi~~ All the weights can be stored without needing to use ~~other~~ & biases

Sparsification / quantization techniques. Thus, accuracy can be maximized.

3

Limitations :

① Latency can not be reduced as the ~~SRAM~~ SRAM & processing units are finite, and operations can only be reordered but not parallelized further.

~~② The computations~~

② In case of large input/output activation maps, ~~the~~ tiling can not be avoided if SRAM is not large enough. This could cause computation overhead.

3. What makes training on tiny devices more challenging than inference? List (1-2 lines each) any two ideas used in MCUNetV3 to enable training on tiny devices. [3]

Training requires calculation of gradients in a directed acyclic graph. For ~~deeper~~ lower layers, more gradients need to be calculated & stored in SRAM which makes training challenging. Also, there is a large loss of accuracy when quantizing from fp32 to int8 as the calculations are different

① MCUNetv3 uses ~~fake~~ scaling of gradients by setting their max & min value to, say -128 to 127, before quantizing. This scaling factor for weights & biases is carried over for calculation & reduces inaccuracy in ~~quantized~~ conv operation after quantization

② During backprop, sparse weight updates are used to update only the most ~~comm~~ important weights (chosen through a heuristic search) & biases (last k layers). ~~An~~

2.5

4. Mention (1-2 lines each) any three approaches used in MCUNet to reduce activation memory requirement for CNNs. [2]

v2

① Tiling of inputs: ~~Across~~

~~Inputs are s~~ sel

The first few layers of ~~input~~ CNNs usually take up the most SRAM memory as the image is of high resolution. The no of input pixels reduces with layers.

For such cases, the ~~smaller inp~~ MCUNetv2 tiles input while reducing SRAM memory & minimizing computation overhead.

↓ into multiple cycles

0.5

②

14

Hardware Architecture for Deep Learning - CS6490. Spring 2024-25.

Dept. of CSE, IIT Hyderabad

Quiz-6

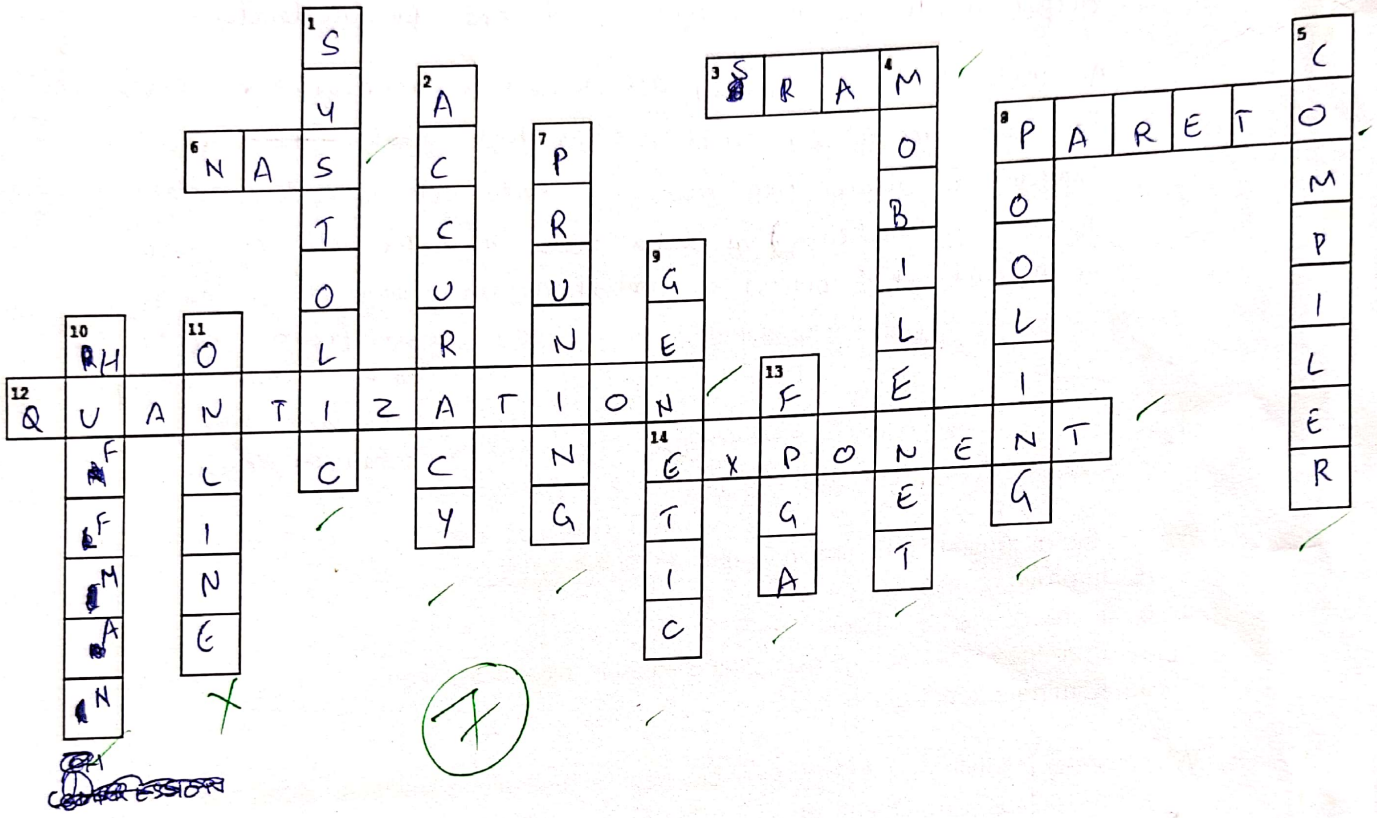
Total marks: 15 **Time: 12 minutes**

Name: P Aditya Sriram

Roll number: EE22BTECH11039

---

1. Solve the given crossword puzzle. [8]



Across

3. The place to store activations in a tiny device
6. Process of identify suitable CNN architecture
8. The non-dominated points during DSE
12. Using INT8 instead of FP32 representation
14. Determines the range of a floating point number

Down

1. An array structure used in most CNN accelerators
2. A common objective metric used in NAS
4. A CNN which proposed depthwise and pointwise convolutions
5. Helps map a CNN graph efficiently to given hardware
7. Removing unimportant weights from a CNN model
8. The layer which helps reduce the input dimensions for next layer
9. A commonly used bio-inspired algorithm for DSE
10. A lossless compression scheme used in ~~Eyeriss~~ Deep compression
11. Accelerators help to reduce this metric
13. A device to implement custom digital logic

2. Briefly explain the importance of using a dual-port SRAM in accelerators. [3]

Accelerators' PEs need to access SRAM to load or read from, as a global memory with the accelerator. It acts as a double buffer b/w accelerator & main (DRAM) memory.

~~The~~ Usually, the weights, ifmaps and ofmaps are loaded/saved to SRAM memory. One port is required for input read and output write to main memory & one for accelerator.

As instructions (high-level) are issued to accelerator (or) in case accelerator card is inserted into ~~memory~~ a system & reads from memory, data gets stored in SRAM (to reduce latency) as data reuse is extremely common in neural nets & CNNs) & simultaneously may be written to or read from the PEs (or their local memory). It acts like a larger "cache".

3

3. Systolic arrays are typically implemented on: [2]
   - a. FPGA
   - b. GPU
   - c. ASIC
   - d. CPU with Vector instructions

2

4. What is the purpose of mutation in Genetic Algorithms? [1]
   - a. Improve crossover
   - b. Increase the population
   - c. Introduce new characteristics not present in initial population
   - d. Eliminate constraints

1

5. Which of the following are examples of bio-inspired heuristic optimization algorithms? [1]
   - a. Particle Swarm Optimization
   - b. Genetic Algorithm
   - c. Gradient Descent
   - d. Ant Colony Optimization

1