Rahul Vigneswaran . K

Hardware Architecture for Deep Learning - CS6490. Spring 2023-24.

Dept. of CSE, IIT Hyderabad

Quiz-5

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

Name: Rahul Vigneswaran K

9

Roll number: CS23MTECH02002

---

1. DNNExplorer generates hardware design: [1]
   [✓ a.] Specific for a given CNN
   b. Common design for many CNNs

1

2. The total number of MAC operations performed by CGNet is always constant for a given CNN: [1]
   a. True
   [✓ b.] False

1

3. How is the pruning in channel gating (CGNet) different from the weight pruning approach? [2]

→ In weight pruning we prune out weights that are below certain threshold in a static manner after training

→ In CGnet we do it dynamically on the fly based on the inputs during the training.

✓ → Weight pruning doesn't depend on input while CGnet does.

→ Channel gating is much more structured wrt input than weight pruning.

→ Channel gating prunes the input effectively reducing comp while weight pruning prunes trained weight.

4. Why does DNNExplorer use a custom pipeline design for each of the few initial layers and the generic structure for the later layers? [3]

→ When checked empirically the variance of CTC for the first half is much higher compared to second half.

→ Higher CTC variance indicates the necessity for much more specialized pipeline than a generic works for all one.

3

5. CGNet uses a banked SRAM structure (splitting the weight values into small sized SRAM banks). Why CGNet needs to use it and how it helps? [3]

→ ~~Doing~~

→ ~~CGNet~~ General conv happens as $W * x$ but CGNet conv happens like $W_p * x_p + W_r * x_r$.

→ Therefore splitting the weight values into small SRAM banks help ~~increase~~ increase the reuse of the banks rather than parsing through the entire weights everytime.

→ CGNet being weight stationary & dynamic in nature will exponentially increase the access cost if not for the banked SRAMs.

2