Assign 2
HADL

Deepshikha CS21BTECH11016

1 TODO

Why 1st layer less than 100?

2 Summary of SCALE-SIM

2.1 What it does

- It is a cycle-accurate simulator for DNN accelerators.
- It takes in the CNN architecture and the accelerator configuration as input and gives the performance metrics as output.
- It computes performance, on-chip and off-chip memory accesss, and interface bandwidth.
- It can implement both scale-up and scale-out instances.

2.2 How it does

- SCALE-SIM generates a cycle-accurate trace of the accelerator execution ,genearting an output which contains SRAM writes.
- THE SRAM trace shows the data movement and computation in the accelerator.
- The requests to SRAM are the DRAM traces , which are used to estimate the interface bandwidth for given CNN.

3 Configs

3.1 Eyeriss

Array size : 12 x 14

3.2 Google

Array size : 256 x 256

3.3 Scale

Array size : 32 x 32

4 Running CNN architecture on SCALE-SIM

4.1 Varying the configs

1. MobileNet
   (a) First layer :
      i. The first conv layer has 224 x 224 x 3 size IFMAP.
      ii. Scale config has 32 x 32 array size , so all of it is utilsed as 224/32 = 7.
      iii. Google config has 256 x 256 array size , it accomodates 224 x 224 x 3 IFMAP , but the rest of the array is wasted and so
      iv. Eyeriss config has 12 x 14 array size , ao it cannot accomodate the entire IFMAP, so 224/12 = 18.66 and 224/14 = 16 , as 224 is not divisible by 12 , so rest of the array is wasted.
   (b) Second Layer :
      i. This is a pooling layer
      ii. Pooling invloves accessing data in a non-regular pattern, which may not fully exploit the regular data access patterns.Thats why,mapping efficiency for all pooling layersis lesser compared to conv layers.
      iii. Eyeriss gives better mapping efficiency than scale because of 112 is divisible by 14.
   (c) Remaining layers :
      - We see that for google config
        – When the number of channels are ¿= 256, mapping efficiency is 100 percent. This is due to the fact that the array size is 256 x 256.
        – As the size of IFMAP is reduced, the mapping efficiency is reduced.
2. Resnet18
   - We see it has no pooling layers, so the mapping efficiency always good.
3.

Varying the architecture

- Eyeriss
  – We see that it gives beter efficiency for resnet18 than mobilenet. The difference is of filters.resnet18 in first layer has 7 x 7 filters while mobilenet has 3 x 3 filters , stride being same for both (2). Since more filter size means more data reuse/parallelism, resnet18 has better efficiency.