Assign 2
HADL

Deepshikha CS21BTECH11016

1 TODO

Why 1st layer less than 100?

2 Summary of SCALE-SIM

2.1 What it does

2.2 How 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.
-

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 CNN architecture

4.1 MobileNet

5 Running CNN architecture on SCALE-SIM

5.1 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.