

Hardware Architecture for Deep Learning - CS6490. Spring 2024-25.
Dept. of CSE, IIT Hyderabad
Assignment-2

The objective of this assignment is to use an estimation tool SCALE-SIM to understand the effect of various configurations on the execution cycles and the DRAM access bandwidth. The suggested steps are some guidelines, but you are free to explore more to develop deeper understanding.

The source paper is:

A. Samajdar, J. M. Joseph, Y. Zhu, P. Whatmough, M. Mattina and T. Krishna, "A Systematic Methodology for Characterizing Scalability of DNN Accelerators using SCALE-Sim," 2020 IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS), Boston, MA, USA, 2020, pp. 58-68. Link: <https://ieeexplore.ieee.org/document/9238602>

1. Download/setup the scale-sim repo as per instructions from <https://github.com/scalesim-project/scale-sim-v2> or <https://scalesim-project.github.io/>
2. Understand what it does, how it does, and write a short summary of your understanding
3. Pickup different CNN architectures (mobilenet, alexnet, resnet18, Googlenet, FasterRCNN, yolo_tiny (pick the yolo_tiny model from [link](#))) and run the tool for three different hardware configs (eyeriss, google, scale) as given in the repo. Tabulate various metrics, study the behavior and summarize the observations.
4. For three CNNs (mobilenet, FasterRCNN, and resnet18), and eyeriss.cfg, study and summarize your observations for the following:
 - a. Change the dataflow architecture between WS, IS, and OS and observe the effect
 - b. Change the size of different SRAM and observe the effect
 - c. Change the array size and observe the effect

Submission:

1. This assignment is to be done individually by all students crediting the course.
2. Audit students are not required to do this assignment.
3. Submit a pdf file containing your answers. Write necessary details only.
4. This should be your own work and not copied from any other source. Please read <https://cse.iith.ac.in/academics/plagiarism-policy.html> for more details.
5. Sufficient time is given, so NO late submission is expected. If submitting late, a 10% penalty for a delay of every 24 hours or part of it.