DAC YF

DNN Training Accelerator

Xingzhou Cheng, BUAA

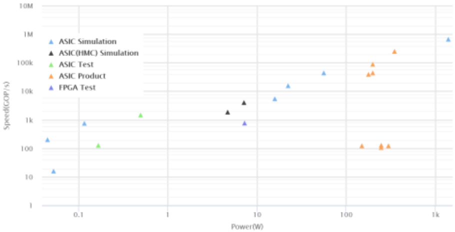
DESIGN AUTOMATION CONFERENCE

Introduction

There has been a lot of work focus on acceleration of DNN inference using special hardware. However , Only a small amount of work is aimed at DNN training. On the one hand, the training process of DNN is generally done offline by GPU,which is insensitive to power consumption and area of hardware. On the other hand, training is much more complicated than inference(inference is just a small step of training),which requires considerable effort even for commercial companies. However , maybe it means more exploration and optimization space for software and hardware designers.

Research status

Much work[1][2] have explored data reuse in the training process.
It has been proved that low-precision training is feasible in CPU or GPU, and some work[3] implemente it with more on special hardware.
Some work [4][5] fine-tune the network for scenes on edge devices, and recently some work [6] has explored the data sparsity in cnn training.



What I can do

Sparse Train (with corase-grained/ fine-grained pruning)

Low-bit / Mixed-precision (maybe fine-grained)

Distributed Training

More data reuse
……

Reference

[1] Y.Chen ,et.al. , MICRO,2014
[2] S.Venkataramani, et.al. , ISCA ,2017
[3] S.Choi, et.al , DAC, 2019
[3] J.Lee, et.al , ISSCC, 2019
[4] Y.Liu ,et.al, JSSC,2019
[6] P.Dai, et.al, DAC,2020