Contents

一、指令选型背景 1. 采用 scratchpad memory 进行数据存储 a. 对于 cnn 网络，卷积运算占大部分算子的运算时间，约为 80%。b. 卷积运算的核心在于矩阵的频繁从内存加载，故采用 scratchpad memory 在片上集成 sram 对于源数据（eg. 卷积核）数据进行暂存，避免大量的访存操作。2. 采用 int8 作为单位元素大小 a. 现有的深度学习框架 比如：TensorFlow，pytorch，Caffe，MixNet 等，在训练一个深度神经网络时，往往都会使用 float 32（Full Precise ，简称 FP32）的数据精度来表示，权值、偏置、激活值等。但是如果一个网络很深的话，比如像 VGG，ResNet 这种，网络参数是极其多的，计算量就更多了（比如 VGG 19.6 billion FLOPS, ResNet-152 11.3 billion FLOPS)。实际上有些人认为，即便在推理时使用低精度的数据（比如 INT8)，在提升速度的同时，也并不会造成太大的精度损失。b. 在现有的许多移动端推理网络如 ncnn 等都是支持 int8 量化的。3. 采用 int8 作为 scratchpad 的单元大小，scratchpad 大小初定为 128*8bit=1Kib a.* 对于 *ncnn* 等网络常用的 *1*1，3*3*，*5*5，7*7 卷积核大小，采用 1Kib 的 scratchpad 大小能够较好的暂存 featuremap 和 kernel 数据。b. scratchpad 从 0 开始索引编址，支持 0-127 个索引的数据读写，每个元素大小为一个字节。