

## 个人信息

- 董启凡/男/1990.08.17
- 手机: 18616265727
- Email: dongqifan1990@gmail.com
- 微信: dongqifanyy



## 教育经历

2009.09-2013.06	本科	山东大学信息科学与工程学院	工学硕士(推免)
2013.09-2016.06	研究生	山东大学信息科学与工程学院	工学学士

## 工作经历

2016.06-至今 酷芯微电子有限公司 (机器视觉与算法工程师)

## 项目经历

### 1. CNN算法的嵌入式移植与优化项目 (2016.12 - 2018.04)

该项目完成了将浮点caffe网络自动移植到DSP上的工作,支持 AlexNet,GoogleNet,ResNet,MobieNet,Yolo等主流网络.主要包括两大部分:

- 1.1 CEVA的Generator的实现  
通过研究CEVA的指令集,搞懂了在DSP端实现CNN的方法以及了加密的Generator端的权重量化方法,重排序方法,并完成了代码实现,产生的文件与CEVA提供的Generator产生的文件一致,主要包括以下几点:
  - 权重系数的重排  
卷积的不同kernel的计算可以并行,受限与CEVA的指令集,需要在进行卷积计算之前,将4个kernel的系数重排使得内存地址连续,提高了内存访问效率
  - 权重系数的量化  
深入研究了权重量化的相关文章,最终基于文章Ristretto的动态固定点方法,实现了generator端的权重系数的量化。包括16bit与8bit权重
  - caffe框架的修改  
generator调用caffelib来解析权重文件并进行forward操作,同时需要保存中间的运算结果以及权重系数来进行量化与重排序操作,并基于该lib完成了各个层参数的保存,最终通过Generator会生成两个二进制文件供给DSP端计算。
- 1.2 CEVA的CDNN算法的DSP端实现
  - CNN算法各个Layer的功能实现
    - convolution layer(包括group,depth wise wise)
    - batchnorm layer(与convolution layer,bias,scale,activation合并计算)
    - leaky-relu,relu等合并计算

- split,concat,element wise等实现
- Pooling(max-pooling,average-pooling,roi-pooling)
- RegionLayer(YOLO),其中使用了expf的快速计算方法,在此基础上完成定点化、向量化
- Layer计算的速度优化
  - 针对128KB DTCM的片上内存的速度优化(最终相对于512KB DTCM仅有低于5%的额外耗时)
    - 新的Tile划分方式 (channel方向划分)
    - 新的调度方式 (充分复用小的DTCM)
    - 新的DMA搬运方式, 减少CPU负载, 提升DMA带宽利用率, 减少了queue memory的尺寸
    - 不同width,height,channel组合下最优计算函数的选取策略
    - 是否按照block划分的修改
  - 针对卷积kernel计算的优化
    - 通过合理设置pattern, 使得没有权重的重排序不需要补0, 将macc效率从75%提升到100%
    - 最大化利用每次load的数据, 减少for循环次数
    - 合理设置stride, 利用好bank parallel,使得1x1的优化效率最高
    - 11x11,7x7,5x5,3x3,1x1等各种情况(stride,512kb DTCM,128kb DTCM)的卷积优化
    - 8bit权重计算优化, 快于ceva
    - 溢出问题的合理规避, 溢出主要有两种, 乘累加的溢出与int->short的溢出。分别从generator端与dsp端进行了合理规避
  - 卷积优化最优的评价标准制定
    - 查看汇编代码看VPU,LSU单元是否全部被并行利用并软件流水
    - 计算macc利用率看是否达到预期
- DMA queue在128KB DTCM的分配
 

硬件仿真的时候, 经常DMA搬运挂掉, 经过dump对应位置的波形并分析, 定位到是不同memory读写latency不一致引起的dma搬运的错乱。CPU在写descriptor的时候, 会去更新queue\_base\_ptr的wptr指针并把descriptor写到memory。如果queue\_base\_ptr位于DTCM而queue位于DDR,会导致wptr++的时候descriptor还没写入DDR.导致queue拿到错误的descriptor。DMA搬运异常
- 硬件仿真profile确定优化瓶颈
 

在veloce仿真平台将运行cnn整个过程的VPU,LSU,DMA等valid信号dump出来,找到VPU利用率低的位置, 查看PC值并根据PC值查看-o4反汇编代码确定源代码运行的位置, 分析VPU利用率低的原因并进行修改, 进一步提升速度, 主要修改有:

  - 由于ping-pong buffer, 在VPU密集计算的时候仍有代码去访问了DDR, 导致CPU wait住。例如const data。
  - 将dma\_queue\_base\_ptr放在片外, push\_desc\_in\_queue会去访问ddr, 导致cpu wait住。
  - matrix计算函数指针放在了DDR,每次tile循环去访问DDR.
  - class的成员变量放在了DDR

- CEVA编译器的bug的定位  
编写的CNN程序软件仿真结果正确，而硬件仿真结果错误。在硬件平台仿真加debug信息或者加trigger信息的时候反而结果会正确。开-o4优化选项结果错误而-o0优化选项结果正确。最终经过debug是因为在Toolbox V16的版本上编译vsspmac的指令与vmax指令时中间会少插一个nop，导致在寄存器的值尚未被写回的时候就被读取。通过dump波形，查看-o4优化选项的汇编代码并人肉定位到源代码的位置，最终确定了问题。

## **2. 扫地机项目的支持 (2017.12 - 2018.03)**

负责扫地机项目的识别算法的训练（前期）与嵌入式移植：

- 算法训练：
  - 搭建了darknet的评估flow。例如map计算，loss曲线绘制
  - 使用imagenet与训练，将该权重作为tiny-yolo网络的初始化权重
  - 数据增强：反转、噪声、random crop等
  - 修改网络结构，保证精度的前提下减少flops
- tiny-yolo网络算法的嵌入式移植
  - 解决了由于芯片没有非线性单元导致的软硬件结果不一致的问题
  - 协同软件组debug内存overlap的问题
  - 将最新训练好的权重，最快部署到芯片，用于demo
  - 自动化浮点与定点评估工具

## **3. 下一代芯片的demo (2018.03 - 2018.04)**

在软件SDK尚未ready的情况下，完成了tiny-yolo算法在裸版上的demo，能够从摄像头采集图像并最终将识别的结果显示在VGA接口的显示器上。该部分主要工作有：

- Baremetal CEVA多核的程序编写以及调试
- 芯片多核worst case的编写以及多核功耗测量
- ARM与CEVA多核通信
- Demo程序的编写以及调试

## **4. Miscellaneous**

- 协助DLA硬件组熟悉CNN算法并协同制定spec
- CEVA-XM4 dsp的培训文档、实验例程与移植guideline整理
- 协助ISP组移植畸变矫正、双目立体视觉匹配算法到CEVA
- 移植人脸检测算法Haar+adaboost算法
- Clang++编译器的研究，直接编译C/C++ codes并与ARM性能比较
- SLAM Eigen库的研究与优化方法探索，协助移植到CEVA
- Vision服务器的搭建
- darknet2caffe转换

## **技能清单**

- 语言水平: CET6
- 编程语言: C/C++,Python,Shell,Halide,Verilog
- 效率工具: GNU Make,CMake,QMake,Markdown,Git
- 开发环境: Visual Studio,Eclipse,PyCharm,Qt Creator,ModelSim,Cl,Gcc/G++,Clang++,LLVM,Gdb
- 计算机视觉: OpenCV,Dlib,Caffe,Tensorflow,Pytorch
- 深度学习网络: AlexNet,GoogLeNet,ResNet,SqueezeNet,MobileNet,ShuffleNet,Faster-RCNN,YOLO,SSD,MTCNN,FCN

## 荣誉与证书

- 2017.12           上海市酷芯微电子公司优秀新人奖
- 2016.05           山东大学2016届优秀毕业生
- 2015.12           山东大学2015年度优秀研究生
- 2015.12           山东大学2015年度光华奖学金
- 2012.10           山东大学优秀学生奖学金
- 2011.10           山东大学优秀学生奖学金、山东大学潍柴动力奖学金

## 文章作品

- Fall Alarm and Inactivity Detection System Design and Implementation on Raspberry Pi  
International Conference on Advanced Communication Technology, **Advanced Communication Technology (ICACT)**, 2015 17th International Conference on
- Sliced integral histogram: an efficient histogram computing algorithm and its FPGA implementation,  
**Multimedia Tools and Applications**, June 2017, Volume 76, Issue 12, pp 14327–14344