SCNN：An Accelerator for Compressed-sparse Convolutional Neural Networks

笔记本：	paper		
创建时间：	2019/4/17 15:57	更新时间：	2019/4/18 0:04
作者：	wangduo		
标签：	DNN-chip		

简介：
SCNN：An Accelerator for Compressed-sparse Convolutional Neural Networks
通过PlanarTiled-InputStationary-CartesianProduct-sparse数据流，同时完成对压缩下权重和数据的复用，映射到新设计的硬件结构完成加速。
ISCA-2017

Angshuman Parashar[†] Minsoo Rhu[†] Anurag Mukkara[‡] Antonio Puglielli[*]
Rangharajan Venkatesan[†] Brucek Khailany[†] Joel Emer[†‡] Stephen W. Keckler[†] William J. Dally[†◇]

NVIDIA[†] Massachusetts Institute of Technology[‡] UC-Berkeley[*] Stanford University[◇]
aparashar@nvidia.com

1.介绍
50%-70%的激活值为0
SCNN压缩激活和权重，激活向量在输入平稳中被重用，在乘法器阵列中同时与一系列权重以笛卡尔积的形式相乘，输出给累加缓存。

2.动机

Table 2: Qualitative comparison of sparse CNN accelerators.

Architecture	Gate MACC	Skip MACC	Skip buffer/ DRAM access	Inner spatial dataflow
Eyeriss [7]	A	–	A	Row Stationary
Cnvlutin [1]	A	A	A	Vector Scalar + Reduction
Cambricon-X [34]	W	W	W	Dot Product
SCNN	A+W	A+W	A+W	Cartesian Product

3.SCNN　数据流

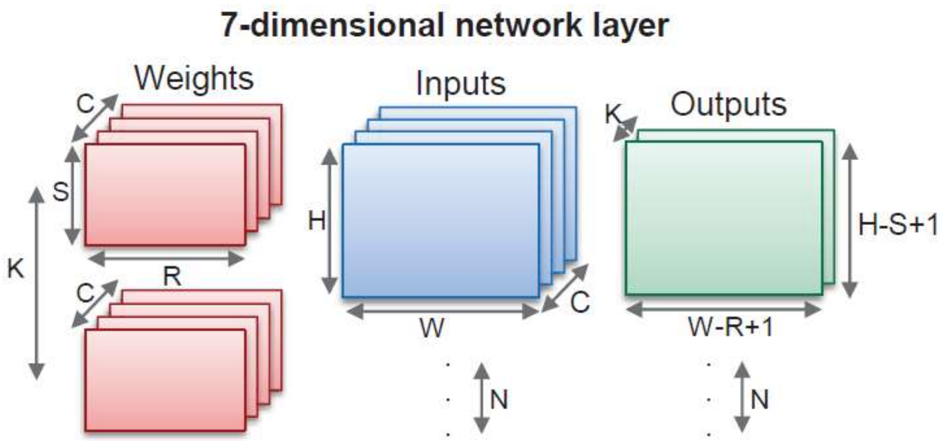


Figure 2: CNN computations and parameters.

```
    BUFFER wt_buf[C][Kc*R*S/F][F];
    BUFFER in_buf[C][Wt*Ht/I][I];
    BUFFER acc_buf[Kc][Wt+R-1][Ht+S-1];
    BUFFER out_buf[K/Kc][Kc*Wt*Ht];
(A) for k' = 0 to K/Kc-1
    {
      for c = 0 to C-1
        for a = 0 to (Wt*Ht/I)-1
        {
(B)       in[0:I-1] = in_buf[c][a][0:I-1];
(C)       for w = 0 to (Kc*R*S/F)-1
          {
(D)         wt[0:F-1] = wt_buf[c][w][0:F-1];
(E)         parallel_for (i = 0 to I-1) x (f = 0 to F-1)
            {
              k = Kcoord(w,f);
              x = Xcoord(a,i,w,f);
              y = Ycoord(a,i,w,f);
(F)           acc_buf[k][x][y] += in[i]*wt[f];
            }
          }
        }
      out_buf[k'][0:Kc*Wt*Ht-1] =
        acc_buf[0:Kc-1][0:Wt-1][0:Ht-1];
    }
```

Figure 4: PT-IS-CP-dense dataflow, single-PE loop nest.

PT-IS-CP-dense 数据流，对权重和激活进行压缩，变为PT-IS-CP-sparse dataflow
PlanarTiled-InputStationary-CartesianProduct-sparse 数据流模式
K/Kc -> C -> W -> H -> Kc -> R -> S
PE内并行：weight x input = F x I
PE间并行：空间切片策略，将工作分配给不同PE独立运行
数据圈问题（边界）：采用输出圈的方式，PE的累加buffer比Kc*Wt*Ht稍大，包含部分不完全的部分和通过PE通信完成计算
将Kc*R*S权重编码进一个block，Wt*Ht激活编码进一个block
传递Wt*Ht区域内的非零激活和其坐标
用压缩格式的非零值坐标进行计算得到输出坐标
累加buffer以非压缩格式索引，通过以交叉开关的scatter网络分布在各个PE中，通过输出索引路由

4.SCNN 加速器架构
针对卷积操作加速
GooLeNet的FC计算占1%

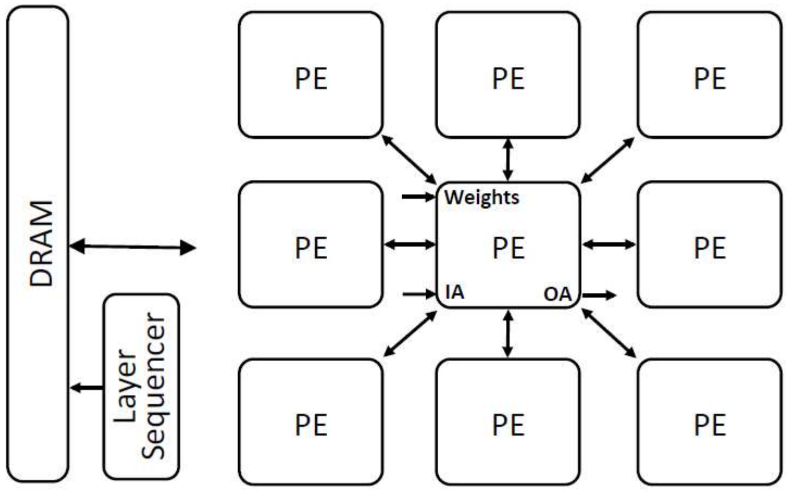


Figure 5: Complete SCNN architecture.

通过Layer Sequencer驱动PE阵列，通过仲裁总线的全局网络实现权重广播，从DRAM的输入激活的点对点传递，和输出激活写回DRAM

8个邻居 PE间传输数据

scatter network传输部分和

IARAM: input activation RAM

IARAM和OARAM可在两层计算序列中逻辑交换，如果一层的输出激活可以作为下一层的输入激活

基于CNN的参数配置控制器

accumulation buffer size A = 2 x F x I 可以足够减少累加bank冲突，双缓冲机制

PPU(post-processing unit):1)PE交换部分和，2)应用非线性激活，pooling和dropout,3)压缩输出激活

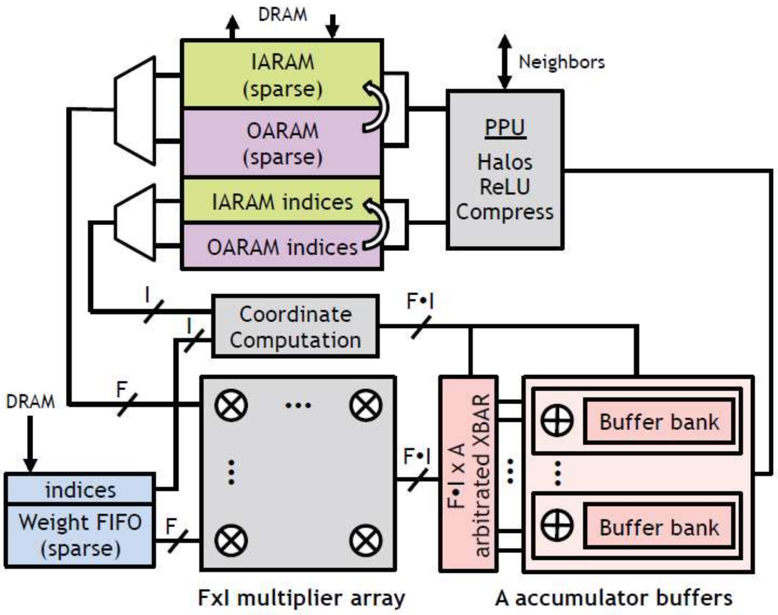


Figure 6: SCNN PE employing the PT-IS-CP-sparse dataflow.

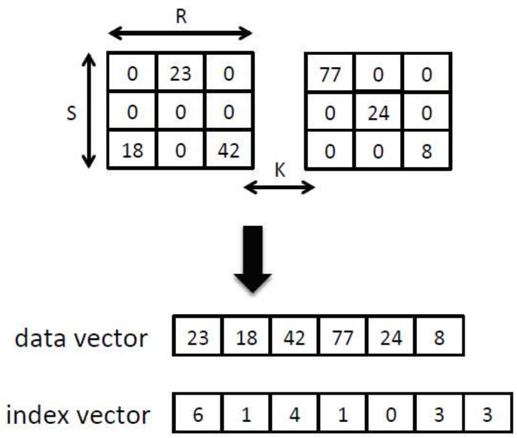


Figure 7: Weight compression.

index vector包含数据数和每个数据之前0的数量
4 bit per index， 最大允许15个零间隔

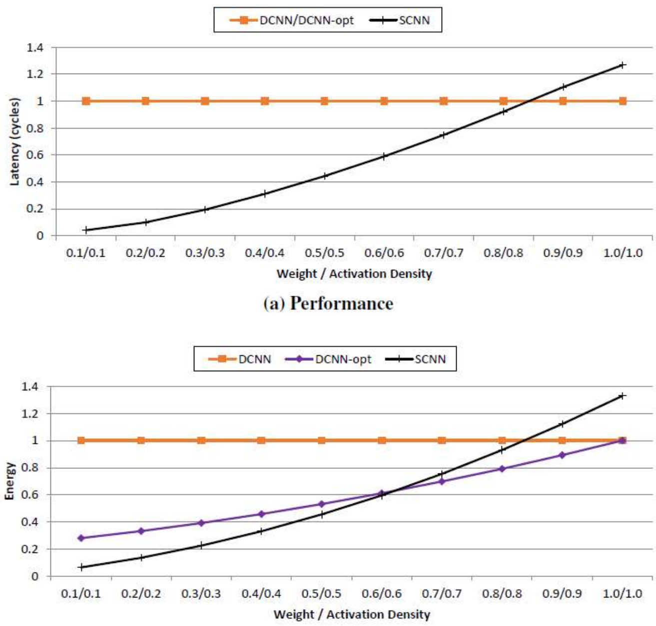
4.4 大模型时间切片
VGGNet中 9 /72 层不能完全填入IARAM\OARAM结构
通过流水线隐藏
激活数据传输的平均18%的每层能耗惩罚

4.5 SCNN架构配置
8x8 PE阵列，每个4*4乘法器阵列，总共1024个乘法器，accumulator buffer 32 bank
PE综合时钟 1GHz, 16nm, 2T-ops峰值通量（16bit mul + 24 bit add）
面积：存储占57%，   乘法器阵列占6%

5.实验方法
周期级别模拟器（pycaffer提取的剪枝权重和稀疏输入激活映射驱动）
TimeLoop 分析模型，探索dense和sparse的设计空间，支持多种数据流
energy model SystemC 实现
1MB的IARAM+OARAM可以支持AlexNet和GooLeNet的激活数据
利用合成网络探索稀疏程度影响

---

6 评估
6.1 CNN稀疏性的敏感性
DCNN-opt:激活值传输给DRAM的压缩和解压缩，乘法门控当输出为0时（节省功耗）



(a) Performance

(b) Energy

Figure 8: GoogLeNet performance and energy versus density.

6.2 SCNN性能和功耗
AlexNet,GoogLeNet,VGGNet的各层：性能，功耗，利用图数据

所有层平均性能：2.37x, 2.19x, 5.52x
影响SCNN与理想状态差距的两个原因：工作集大小和负载均衡(层同步barrier)
平均能耗：2.3x

6.3 PE粒度
跨PE全局barrier vs PE内乘法阵列碎片
8x8 PE,16 mul per PE -> 2x2 PE, 256 mul per PE：11%加速

6.4 权重和激活稀疏性的影响
SCNN-SparseA -> Cnvlutin
SCNN-SparseW -> Cambricon-X

Table 6: Characteristics of evaluated accelerators.

Architecture	Gate MACC	Skip MACC	Skip Buffer Access	Skip DRAM Access	Inner Spatial Dataflow
DCNN	–	–	–	–	Dot Product
DCNN-opt	A+W	–	–	A+W	Dot Product
SCNN-SparseA	A	A	A	A	Cartesian Product
SCNN-SparseW	W	W	W	W	Cartesian Product
SCNN	A+W	A+W	A+W	A+W	Cartesian Product

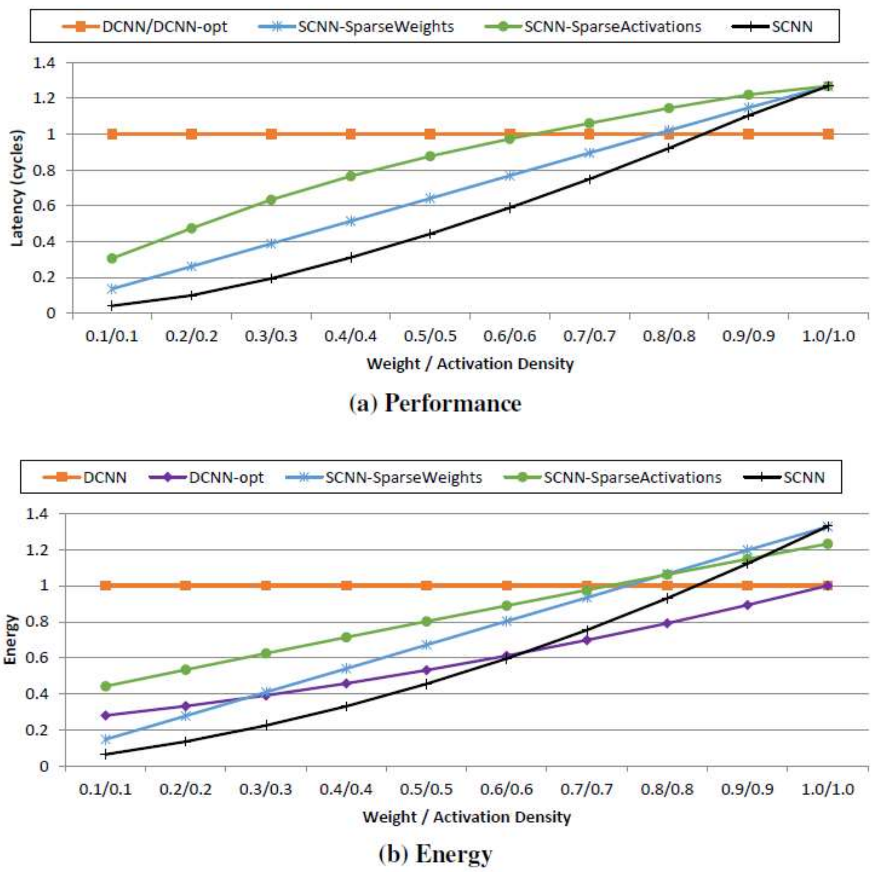


Figure 12: GoogLeNet performance and energy versus density for sparse weights, sparse activations, and both

正常0.4/0.4: 比SparseW和SparseA：性能1.7x/2.6, 功耗1.6x/2.1x

---

7 相关工作
Eyeriss:
只旁路0权重，不剪枝的网络权重稀疏性差，省功耗但不省时间
数据传输：运行长度编码
Cnvlutin:

压缩激活值，不剪枝利用权重稀疏
Cambricon-X:
权重稀疏，在buffer只保留非零权重，不压缩激活和跳过0激活计算
DLAC：只提到0值跳过，没有使用
EIE：
权重和激活压缩，只送0激活，为FC设计
Fused Layer CNN Accelerator (Micro-2016)
混淆邻接层，片上保存中间激活