EIE: Efficient Inference Engine on Compressed Deep Neural Network

EIE：压缩深度神经网络的高效推理引擎

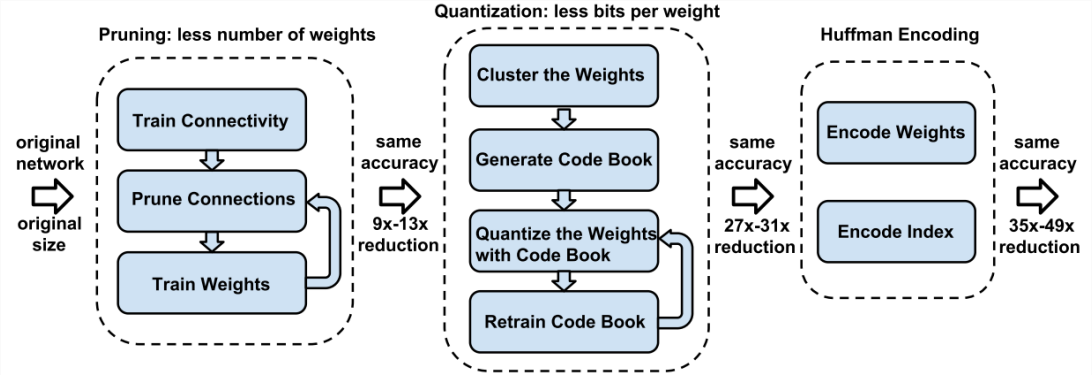
ISCA2016

汇报人：王铎【中国科学院大学】

May 15, 2019

目录

Deep Compression



AlexNet 35X，VGG-16 49X
ICLR-2016 best paper

目录

问题

之前的工作集中在紧密无压缩模型

论文关注点

压缩矩阵乘法， FC层

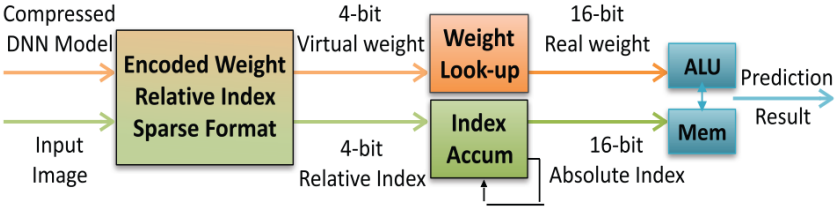


Figure 1. Efficient inference engine that works on the compressed deep

目录

解决方案-PE工作分配

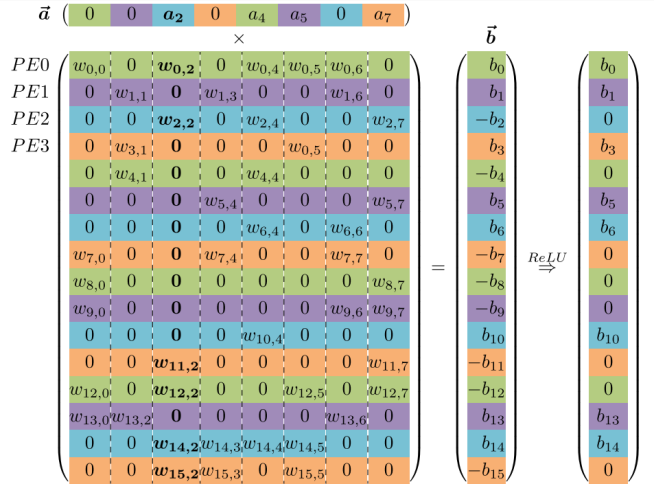


Figure 2. Matrix W and vectors a and b are interleaved over 4 PEs. Elements of the same color are stored in the same PE.

解决方案-CSC压缩格式

Virtual Weight	$W_{0,0}$	$W_{8,0}$	$W_{12,0}$	$W_{4,1}$	$W_{0,2}$	$W_{12,2}$	$W_{0,4}$	$W_{4,4}$	$W_{0,5}$	$W_{12,5}$	$W_{0,6}$	$W_{8,7}$	$W_{12,7}$
Relative Row Index	0	1	0	1	0	2	0	0	0	2	0	2	0
Column Pointer	0	3	4	6	6	8	10	11	13				

Figure 3. Memory layout for the relative indexed, indirect weighted and interleaved CSC format, corresponding to PE_0 in Figure 2.

virtual weight(4b), row index(4b), pointer(16b)

解决方案-架构

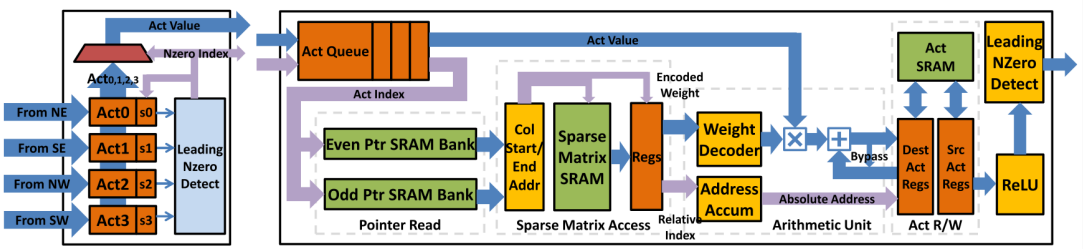


Figure 4. (a) The architecture of Leading Non-zero Detection Node. (b) The architecture of Processing Element.

非零a_j通过quadtree收集，再通过H-tree广播(利用了稀疏性)

Central Control Unit(CCU)：I/O mode + Computing mode

目录

实验结果

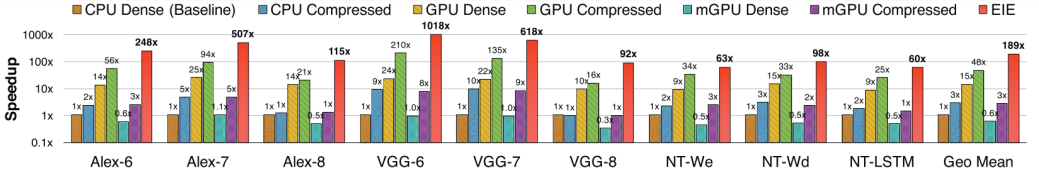


Figure 6. Speedups of GPU, mobile GPU and EIE compared with CPU running uncompressed DNN model. There is no batching in all cases.

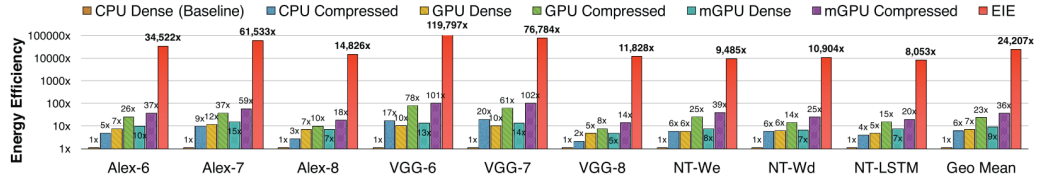


Figure 7. Energy efficiency of GPU, mobile GPU and EIE compared with CPU running uncompressed DNN model. There is no batching in all cases.

References I

[1] S. Han, H. Mao, and W. J. Dally.
Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding.
***Fiber*, 56(4):3–7, 2015.**

Thanks

EIE: Efficient Inference Engine on Compressed Deep Neural Network

EIE：压缩深度神经网络的高效推理引擎