

# 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*

## 1 介绍

---

- 背景
- 问题

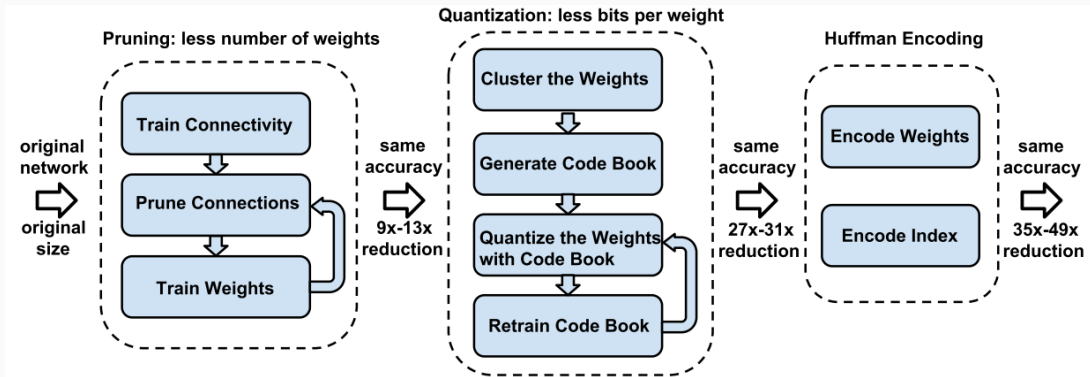
## 2 解决方案

---

## 3 实验

---

# Deep Compression



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

## 1 介绍

---

- 背景
- 问题

## 2 解决方案

---

## 3 实验

---

## 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

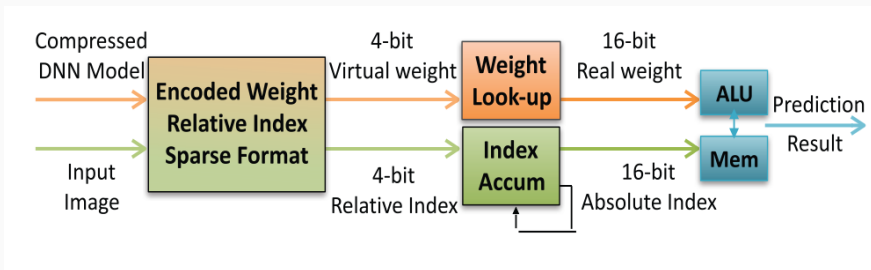


Figure 1. Efficient inference engine that works on the compressed deep

## 1 介绍

---

## 2 解决方案

---

- 解决方案

## 3 实验

---

# 解决方案-PE工作分配

$$\begin{array}{c}
 \vec{a} \begin{pmatrix} 0 & 0 & a_2 & 0 & a_4 & a_5 & 0 & a_7 \end{pmatrix} \\
 \times \\
 \begin{array}{c}
 PE0 \\ PE1 \\ PE2 \\ PE3
 \end{array}
 \begin{pmatrix}
 w_{0,0} & 0 & w_{0,2} & 0 & w_{0,4} & w_{0,5} & w_{0,6} & 0 \\
 0 & w_{1,1} & 0 & w_{1,3} & 0 & 0 & w_{1,6} & 0 \\
 0 & 0 & w_{2,2} & 0 & w_{2,4} & 0 & 0 & w_{2,7} \\
 0 & w_{3,1} & 0 & 0 & 0 & w_{0,5} & 0 & 0 \\
 0 & w_{4,1} & 0 & 0 & w_{4,4} & 0 & 0 & 0 \\
 0 & 0 & 0 & w_{5,4} & 0 & 0 & 0 & w_{5,7} \\
 0 & 0 & 0 & 0 & w_{6,4} & 0 & w_{6,6} & 0 \\
 w_{7,0} & 0 & 0 & w_{7,4} & 0 & 0 & w_{7,7} & 0 \\
 w_{8,0} & 0 & 0 & 0 & 0 & 0 & 0 & w_{8,7} \\
 w_{9,0} & 0 & 0 & 0 & 0 & 0 & w_{9,6} & w_{9,7} \\
 0 & 0 & 0 & 0 & w_{10,4} & 0 & 0 & 0 \\
 0 & 0 & w_{11,2} & 0 & 0 & 0 & 0 & w_{11,7} \\
 w_{12,0} & 0 & w_{12,2} & 0 & 0 & w_{12,5} & 0 & w_{12,7} \\
 w_{13,0} & w_{13,2} & 0 & 0 & 0 & 0 & w_{13,6} & 0 \\
 0 & 0 & w_{14,2} & w_{14,3} & w_{14,4} & w_{14,5} & 0 & 0 \\
 0 & 0 & w_{15,2} & w_{15,3} & 0 & w_{15,5} & 0 & 0
 \end{pmatrix}
 \end{array}
 =
 \begin{array}{c}
 \vec{b} \\
 \begin{pmatrix}
 b_0 \\ b_1 \\ -b_2 \\ b_3 \\ -b_4 \\ b_5 \\ b_6 \\ -b_7 \\ -b_8 \\ -b_9 \\ b_{10} \\ -b_{11} \\ -b_{12} \\ b_{13} \\ b_{14} \\ -b_{15}
 \end{pmatrix}
 \end{array}
 \xRightarrow{ReLU}
 \begin{pmatrix}
 b_0 \\ b_1 \\ 0 \\ b_3 \\ 0 \\ b_5 \\ b_6 \\ 0 \\ 0 \\ 0 \\ b_{10} \\ 0 \\ 0 \\ b_{13} \\ b_{14} \\ 0
 \end{pmatrix}$$

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.

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)



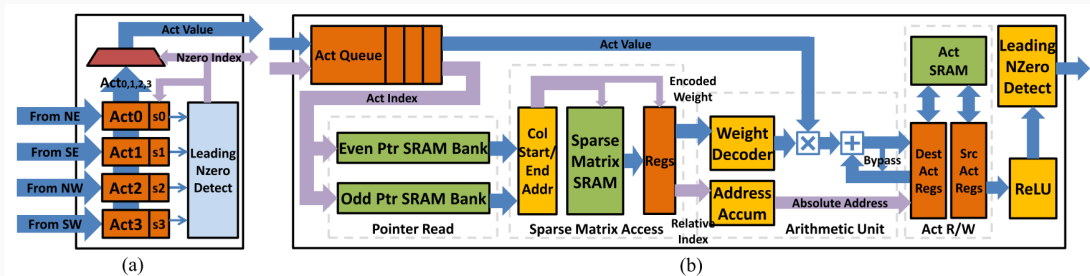


Figure 4. (a) The architecture of Leading Non-zero Detection Node. (b) The architecture of Processing Element.

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

## 1 介绍

---

## 2 解决方案

---

## 3 实验

---

- 实验

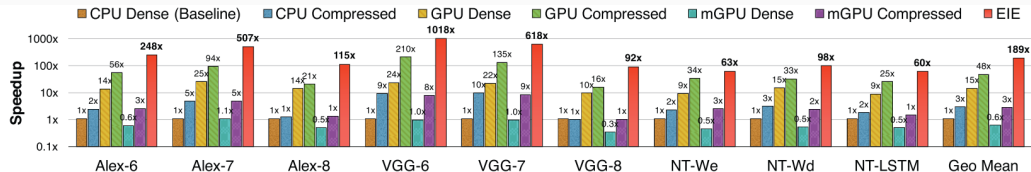


Figure 6. Speedups of GPU, mobile GPU and EIE compared with CPU running uncompressed DNN model. There is no batching in all cases.

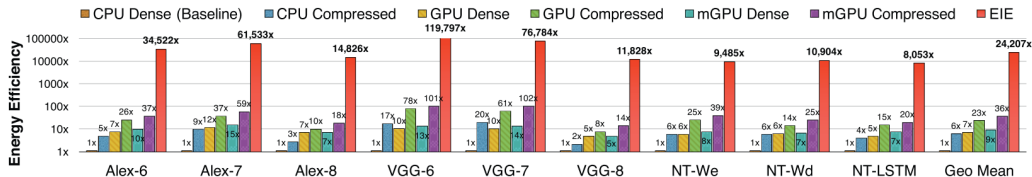


Figure 7. Energy efficiency of GPU, mobile GPU and EIE compared with CPU running uncompressed DNN model. There is no batching in all cases.

[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