EIE: Efficient Inference Engine on Compressed Deep Neural Network

Song Han, Xingyu Liu, Huizi Mao, Jing Pu, Ardavan Pedram, Mark Horowitz, Bill Dally

Stanford University

Jun 20 2016

Presented by Sihang Liu

Slides mostly from Han's presentation

Background: Hardwares for Deep Neural Network

CPU

Background: Hardwares for Deep Neural Network

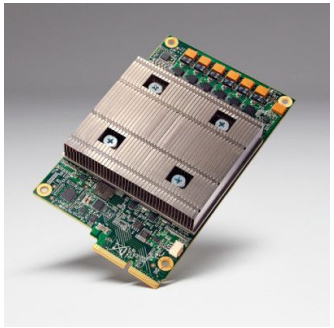
CPU

GPU

Background: Hardwares for Deep Neural Network

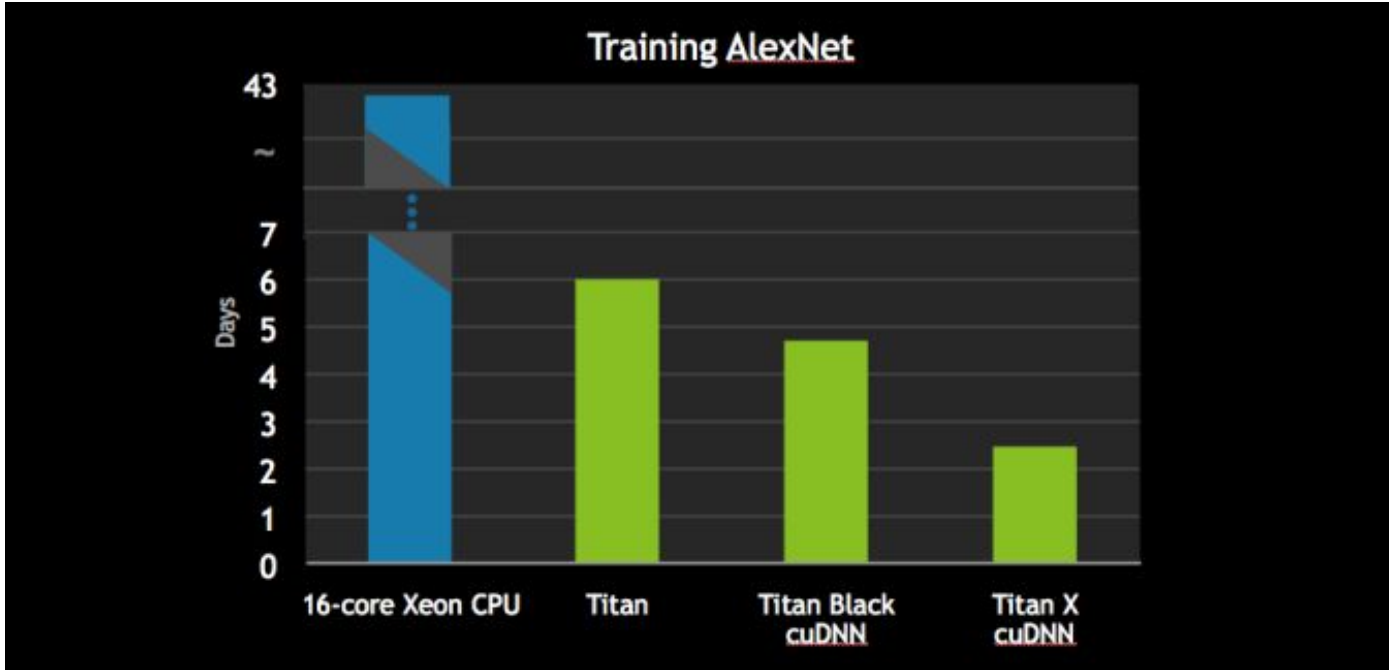
CPU

GPU



Tensor Processing Unit (TPU)

Background: Speedup from GPU



Source: nvidia.com

Deep Learning on Mobile

Phones

Drones

Robots

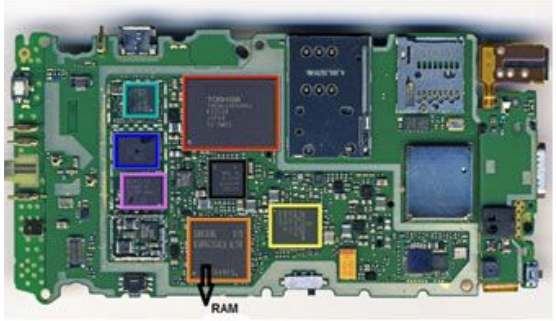
Glasses

Self Driving Cars

Battery Constrained!

Difficulty?

Model Size!

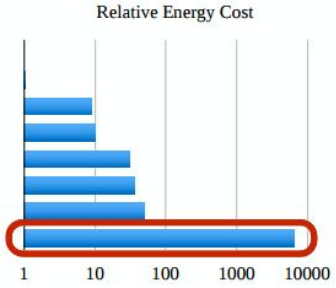


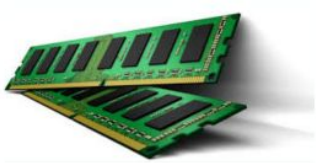
Motherboard of a smart phone

Difficulty?

Model Size!

Operation	Energy [pJ]	Relative Cost
32 bit int ADD	0.1	1
32 bit float ADD	0.9	9
32 bit Register File	1	10
32 bit int MULT	3.1	31
32 bit float MULT	3.7	37
32 bit SRAM Cache	5	50
32 bit DRAM Memory	**640**	**6400**



1  = 100 ✕ +

Deep Compression

Problem: DNN model too large

Solution: Deep Compression

Smaller Size

90% zeros in weights
4-bit weight

Accuracy

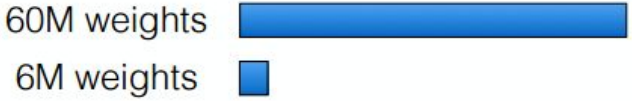
No loss of accuracy /
Improved accuracy

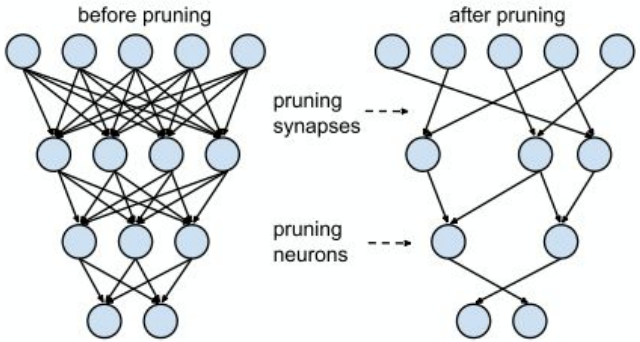
On-chip

State-of-the-art DNN
fit on-chip SRAM

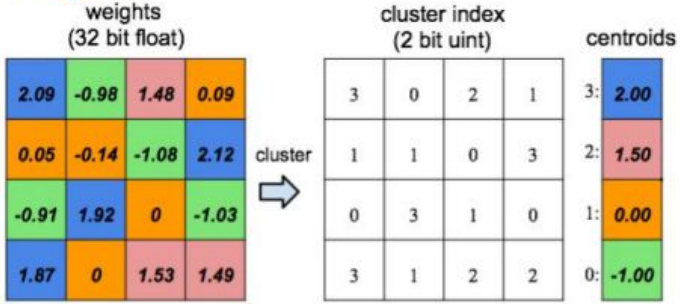
Deep Compression (cont.)

- **Network Pruning [1]**: 10x fewer weights



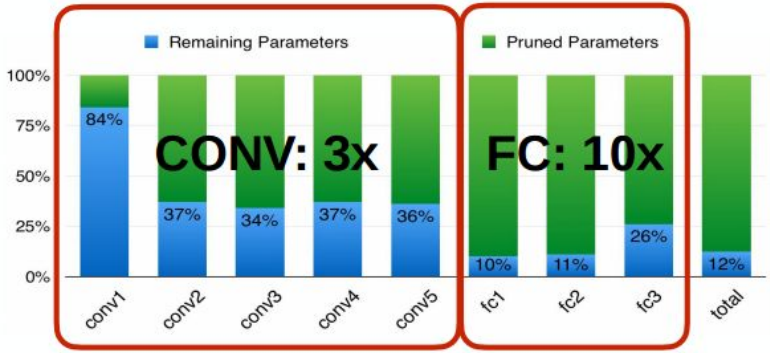


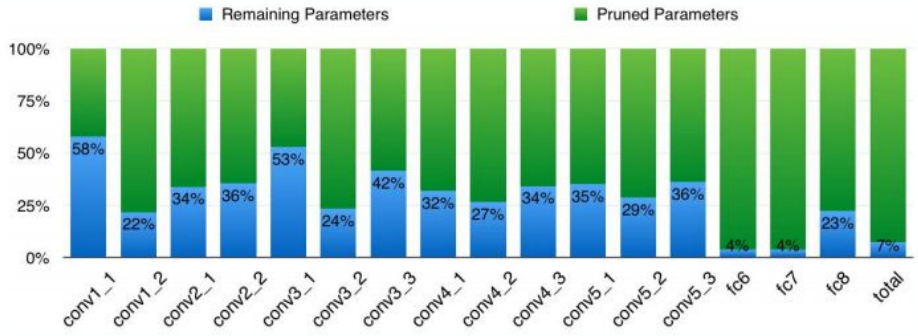
- **Weight Sharing [2]**: only 4-bits per remaining weight



[1]. Han et al. NIPS 2015
[2]. Han et al. ICLR 2016, best paper award

Pruning AlexNet & VGGNet





AlexNet

Layer	Weights	FLOP	Act%	Weights%	FLOP%
conv1	35K	211M	88%	84%	84%
conv2	307K	448M	52%	38%	33%
conv3	885K	299M	37%	35%	18%
conv4	663K	224M	40%	37%	14%
conv5	442K	150M	34%	37%	14%
fc1	38M	75M	36%	9%	3%
fc2	17M	34M	40%	9%	3%
fc3	4M	8M	100%	25%	10%
Total	61M	1.5B	54%	**11%**	**30%**

VGG-16

Layer	Weights	FLOP	Act%	Weights%	FLOP%
conv1_1	2K	0.2B	53%	58%	58%
conv1_2	37K	3.7B	89%	22%	12%
conv2_1	74K	1.8B	80%	34%	30%
conv2_2	148K	3.7B	81%	36%	29%
conv3_1	295K	1.8B	68%	53%	43%
conv3_2	590K	3.7B	70%	24%	16%
conv3_3	590K	3.7B	64%	42%	29%
conv4_1	1M	1.8B	51%	32%	21%
conv4_2	2M	3.7B	45%	27%	14%
conv4_3	2M	3.7B	34%	34%	15%
conv5_1	2M	925M	32%	35%	12%
conv5_2	2M	925M	29%	29%	9%
conv5_3	2M	925M	19%	36%	11%
fc6	103M	206M	38%	4%	1%
fc7	17M	34M	42%	4%	2%
fc8	4M	8M	100%	23%	9%
total	138M	30.9B	64%	**7.5%**	**21%**

Deep Compression (cont.)

Network	Original Size	Compressed Size	Compression Ratio	Original Accuracy	Compressed Accuracy
AlexNet	240MB	6.9MB	**35x**	80.27%	80.30%
VGGNet	550MB	11.3MB	**49x**	88.68%	89.09%
GoogleNet	28MB	2.8MB	**10x**	88.90%	88.92%
SqueezeNet	4.8MB	0.47MB	**10x**	80.32%	80.35%

- No loss of accuracy
- 120X less energy

Accelerator for Compressed Sparse Neural Network

Problem: Irregular Computation Pattern

Solution: EIE accelerator

Sparse Matrix

90% *static* sparsity in the weights,
10x less computation,
5x less memory footprint

Sparse Vector

70% *dynamic* sparsity in the activation
3x less computation

Weight Sharing

4bits weights
8x less memory footprint

Fully fits in SRAM

120x less energy than DRAM

Distributed Storage and Processing

logically

$$\vec{a}\ \begin{pmatrix} 0 & \boldsymbol{a_1} & 0 & a_3 \end{pmatrix}$$

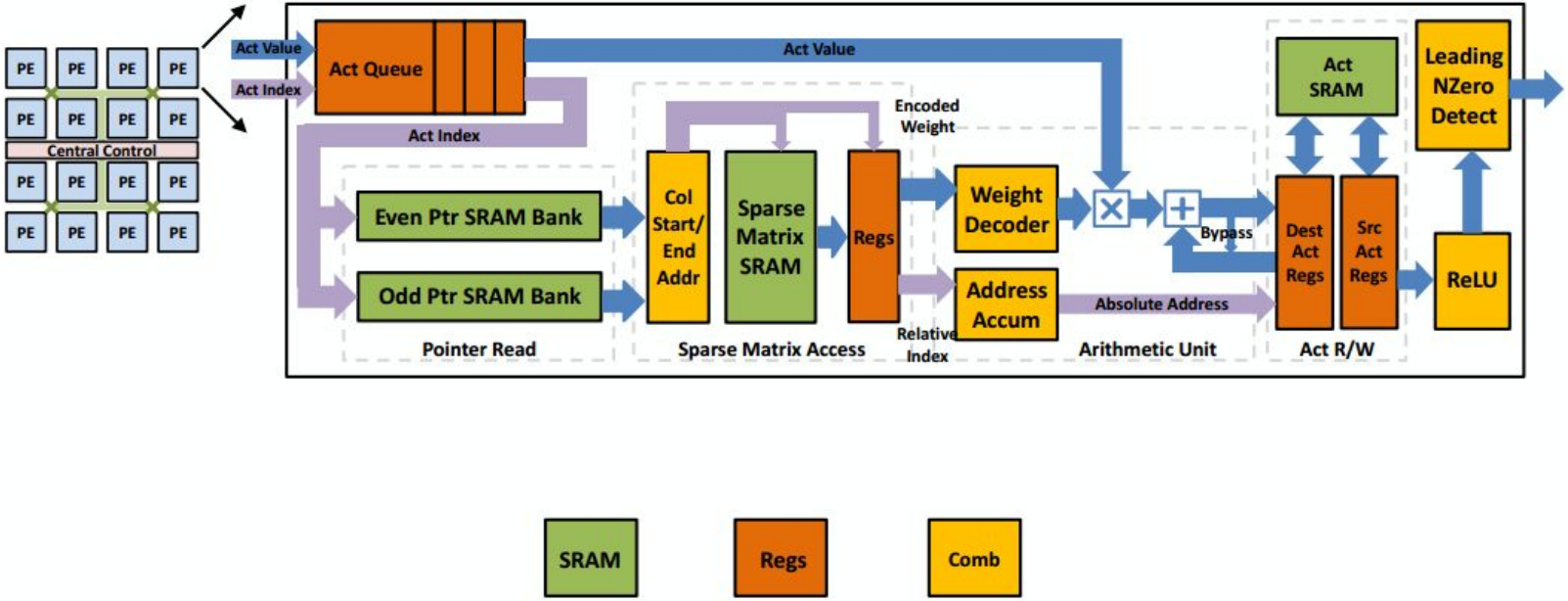
$$\times$$

$$\begin{array}{c} PE0 \\ PE1 \\ PE2 \\ PE3 \\ \\ \\ \\ \\ \end{array} \begin{pmatrix} w_{0,0} & \boldsymbol{w_{0,1}} & 0 & w_{0,3} \\ 0 & \boldsymbol{0} & w_{1,2} & 0 \\ 0 & \boldsymbol{w_{2,1}} & 0 & w_{2,3} \\ 0 & \boldsymbol{0} & 0 & 0 \\ 0 & \boldsymbol{0} & w_{4,2} & w_{4,3} \\ w_{5,0} & \boldsymbol{0} & 0 & 0 \\ 0 & \boldsymbol{0} & 0 & w_{6,3} \\ 0 & \boldsymbol{w_{7,1}} & 0 & 0 \end{pmatrix} = \begin{pmatrix} b_0 \\ b_1 \\ -b_2 \\ b_3 \\ -b_4 \\ b_5 \\ b_6 \\ -b_7 \end{pmatrix} \overset{ReLU}{\Rightarrow} \overset{\vec{b}}{\begin{pmatrix} b_0 \\ b_1 \\ 0 \\ b_3 \\ 0 \\ b_5 \\ b_6 \\ 0 \end{pmatrix}}$$

physically

Virtual Weight	$W_{0,0}$	$W_{0,1}$	$W_{4,2}$	$W_{0,3}$	$W_{4,3}$
Relative Index	0	1	2	0	0
Column Pointer	0	1	2	3	

PE Architecture

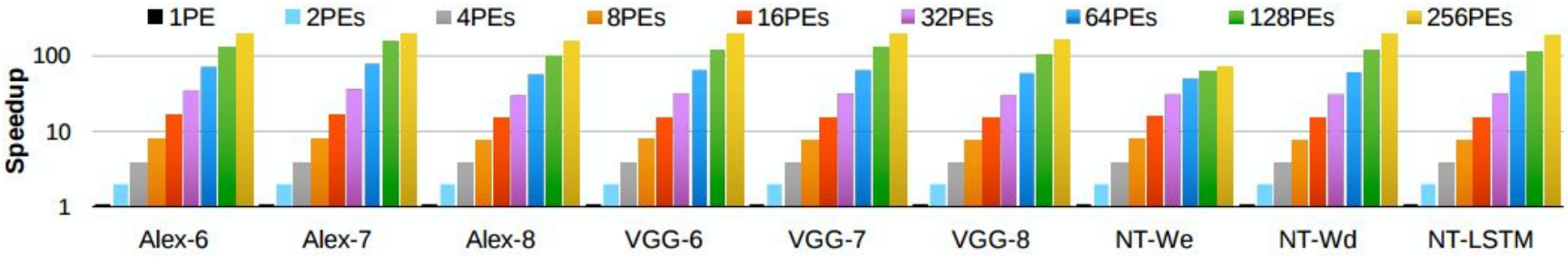


Benchmark

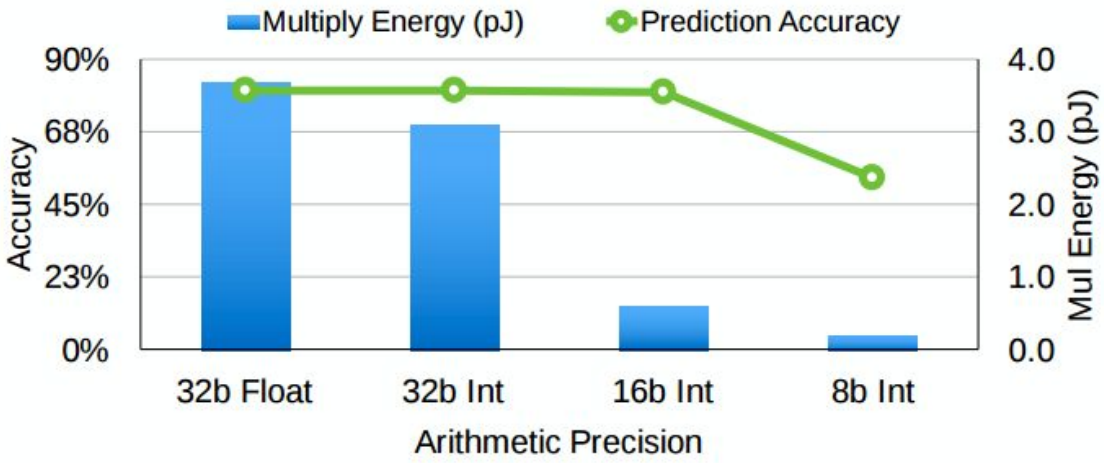
- CPU: Intel Core-i7 5930k
- GPU: NVIDIA Titan X
- Mobile GPU: NVIDIA Jetson TK1

Layer	Size	Weight Density	Activation Density	FLOP %	Description
AlexNet-6	4096 × 9216	9%	35.1%	3%	AlexNet for image classification
AlexNet-7	4096 × 4096	9%	35.3%	3%	
AlexNet-8	1000 × 4096	25%	37.5%	10%	
VGG-6	4096 × 25088	4%	18.3%	1%	VGG-16 for image classification
VGG-7	4096 × 4096	4%	37.5%	2%	
VGG-8	1000 × 4096	23%	41.1%	9%	
NeuralTalk-We	600 × 4096	10%	100%	10%	RNN and LSTM for image caption
NeuralTalk-Wd	8791 × 600	11%	100%	11%	
NeuralTalk-LSTM	2400 × 1201	10%	100%	11%	

Scalability



Prediction Accuracy



Comparison: Throughput