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
June 20, 2016

Deep Learning on Mobile

Phones

Drones

Robots

Glasses

Self Driving Cars

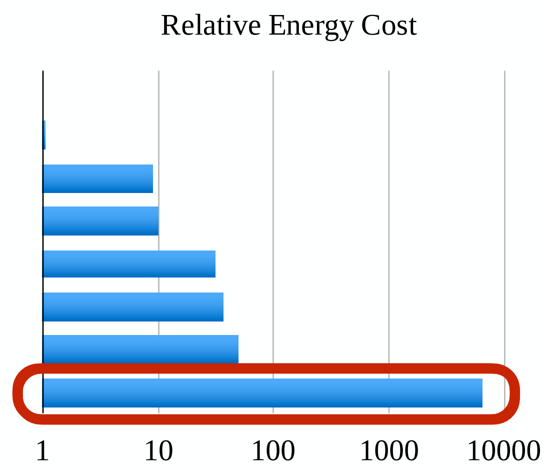
Battery Constrained!

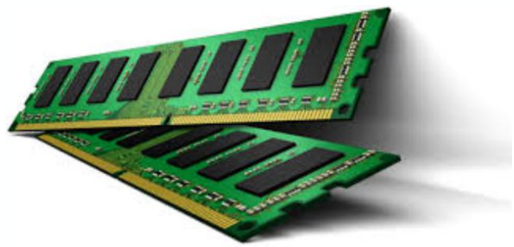
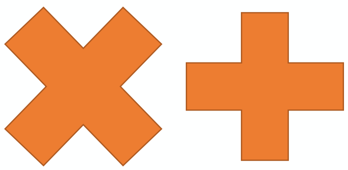
Deep Learning on Mobile: Difficulty?

Model Size!

Accurate Prediction => Large Model => More Memory Reference => High Power

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 

Our Past Work: Deep Compression

Problem 1: DNN Model Size too Large
Solution 1: Deep Compression

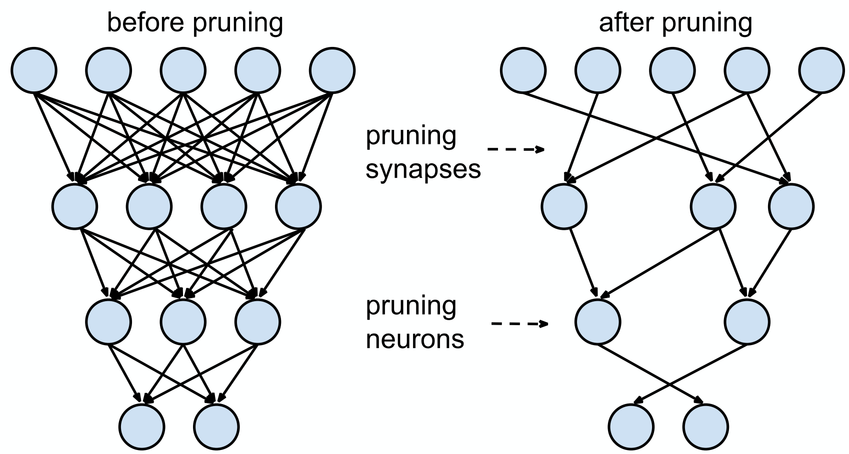
Smaller Size
90% zeros in weights
4-bit weight

Accuracy
No loss of accuracy /
Improved accuracy

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

Our Past Work: Deep Compression

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



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

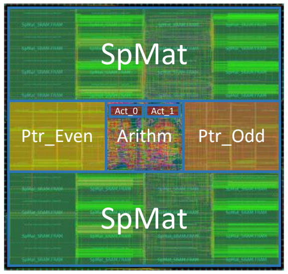
[1]. Han et al. NIPS 2015
[2]. Han et al. ICLR 2016, best paper award

Deep Compression Results

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 on ImageNet dataset.
- Weights fits on-chip SRAM, taking 120x less energy than DRAM.

EIE: First Accelerator for Compressed Sparse Neural Network



Problem 2: Irregular Computation Pattern

Solution 2: 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

Savings are **multiplicative**: 5x3x8x120=14,400 theoretical energy improvement.

Distributed Storage and Processing

logically

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

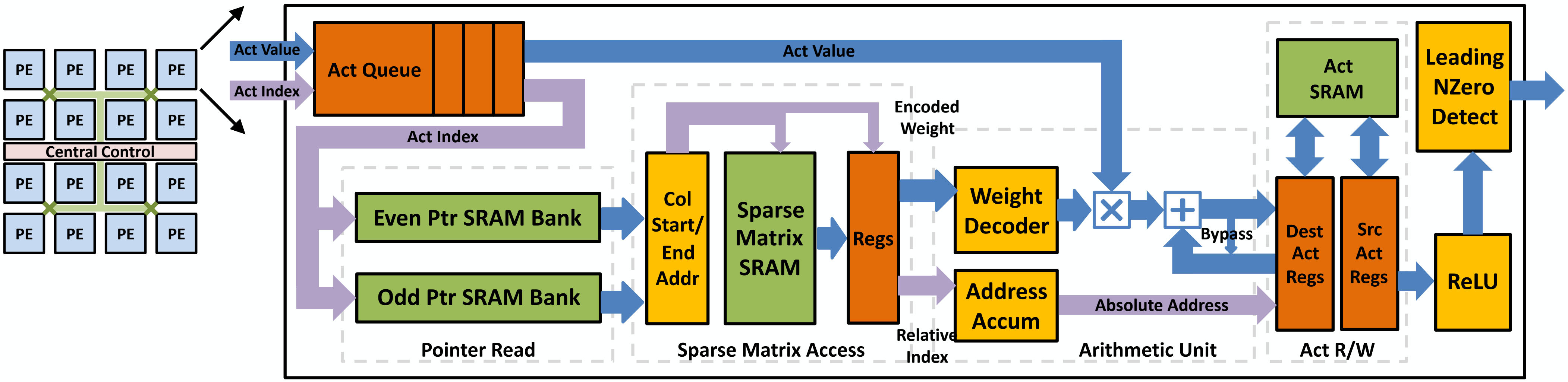
$$\times$$

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

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

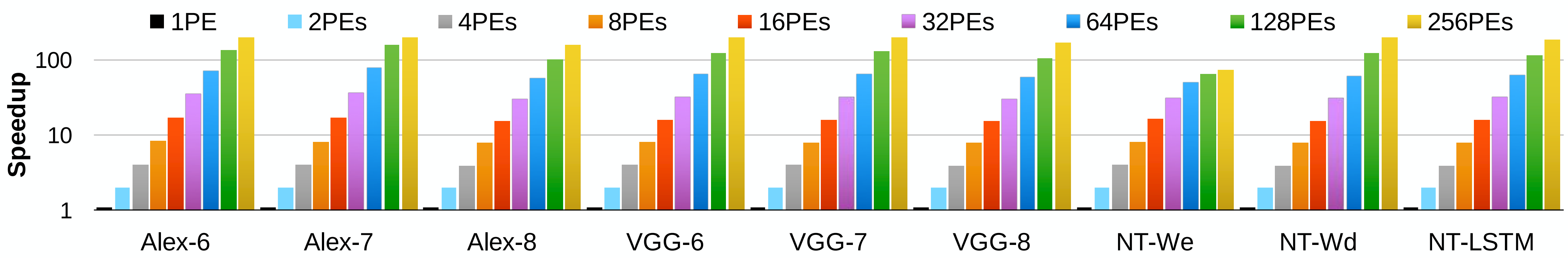


Benchmark

- CPU: Intel Core-i7 5930k
- GPU: NVIDIA TitanX
- 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%	

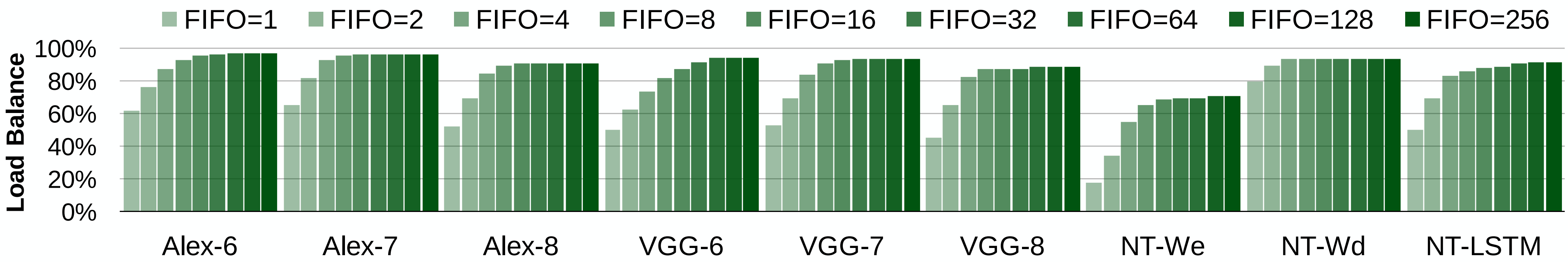
Scalability



#PEs ~ Speedup

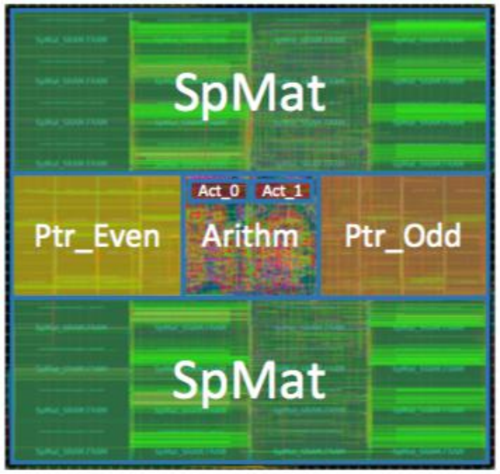
- 64PEs: 64x
- 128PEs: 124x
- 256PEs: 210x

Load Balancing



- Imbalanced non-zeros among PEs degrades system utilization.
- This load imbalance could be solved by FIFO.
- With FIFO depth=16, ALU utilization is > 90%.

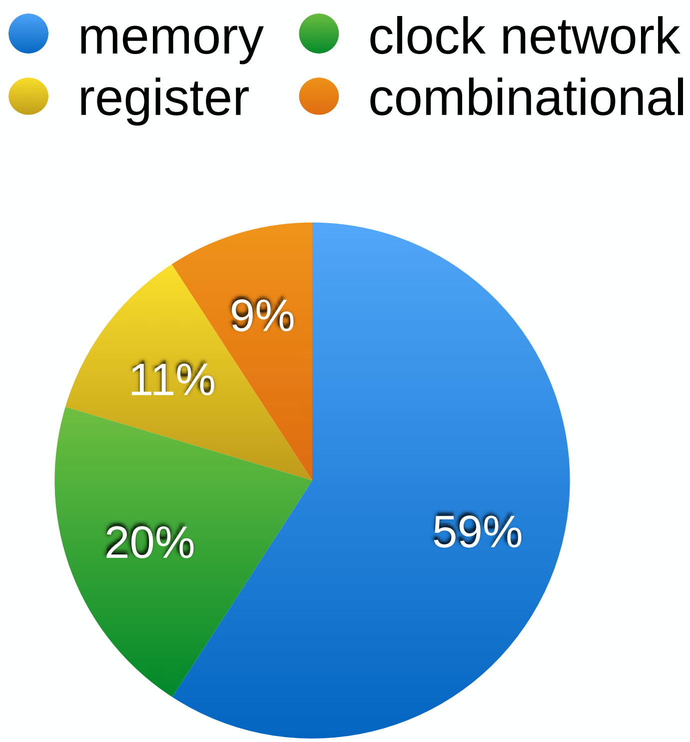
Result of EIE

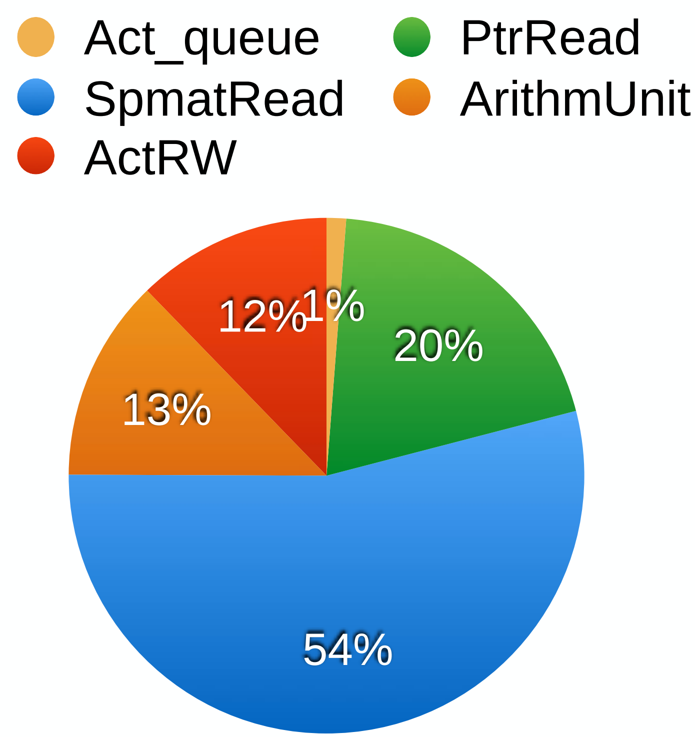


Technology	45 nm
# PEs	64
on-chip SRAM	8 MB
Max Model Size	84 Million
Static Sparsity	10x
Dynamic Sparsity	3x
Quantization	4-bit
ALU Width	16-bit
Area	40.8 mm^2
MxV Throughput	81,967 layers/s
Power	586 mW

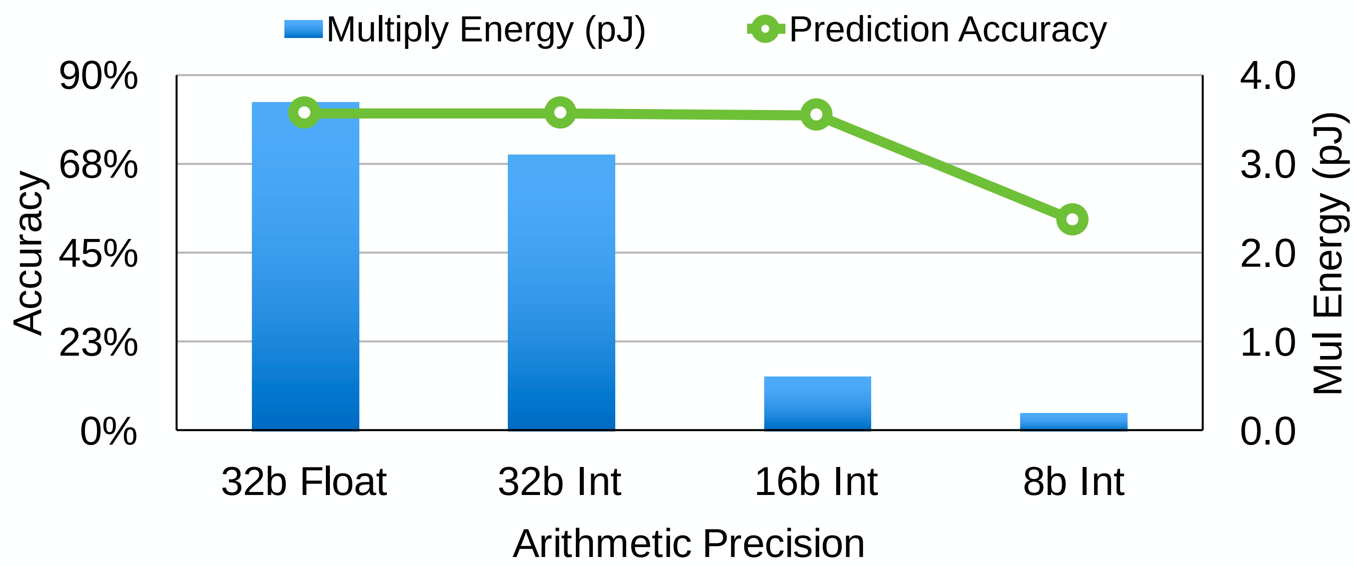
1. **Post layout result**
2. **Throughput measured on AlexNet FC-7**

Energy Breakdown





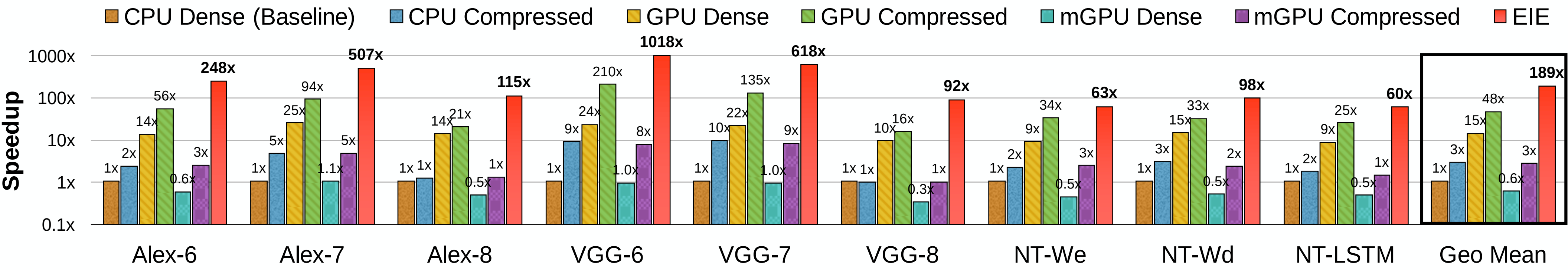
Prediction Accuracy



Mixed Precision:

- 4 bit index (virtual weight)
- 16 bit real weight, 16 bit fixed point ALU

FC Layer: Speedup on EIE



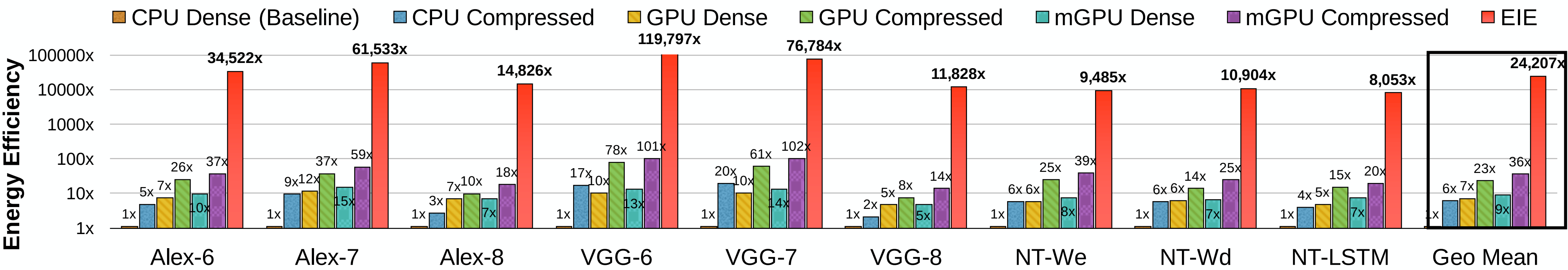
Compared to CPU and GPU:

189x and 13x faster

Baseline:

- Intel Core i7 5930K: MKL CBLAS GEMV, MKL SPBLAS CSRMV
- NVIDIA GeForce GTX Titan X: cuBLAS GEMV, cuSPARSE CSRMV
- NVIDIA Tegra K1: cuBLAS GEMV, cuSPARSE CSRMV

FC Layer: Energy Efficiency on EIE



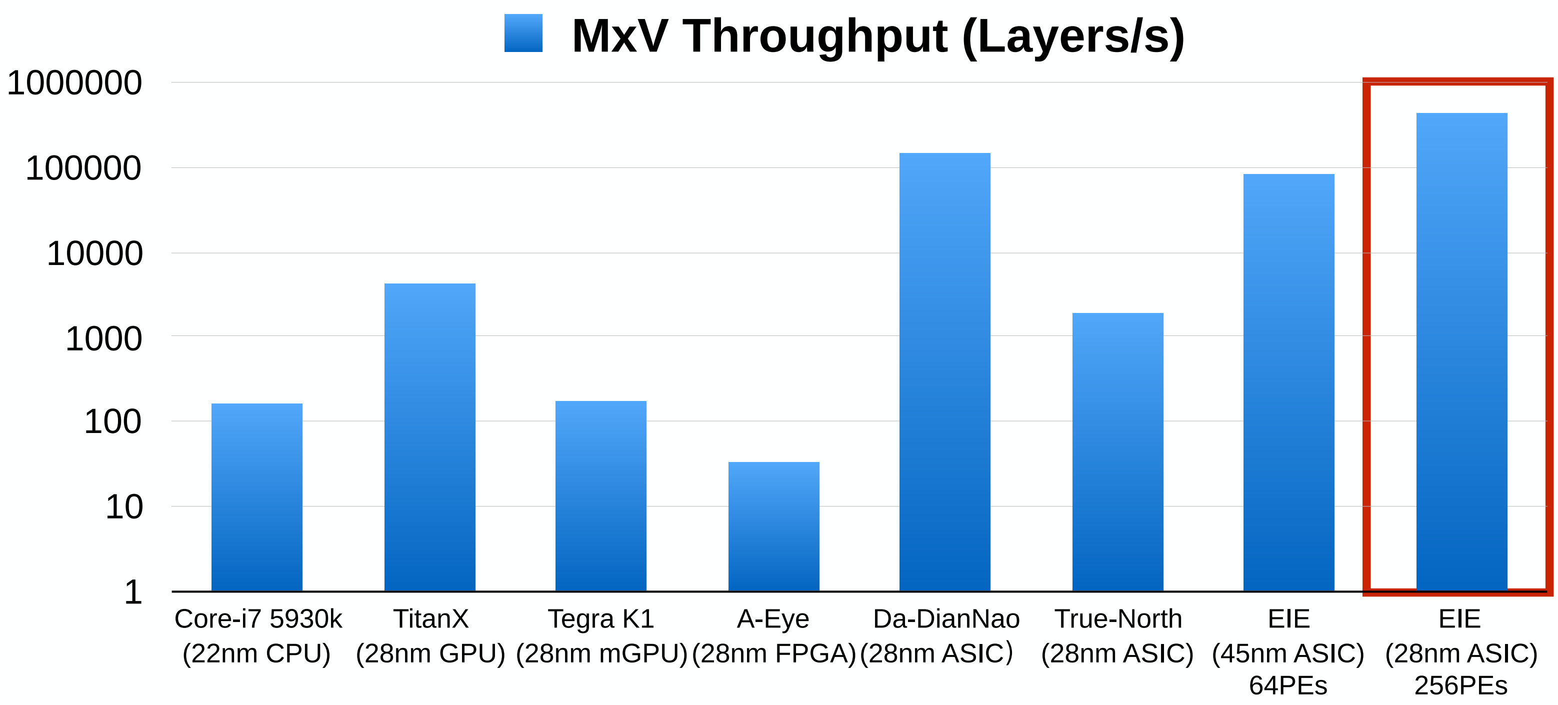
Compared to CPU and GPU:

24,000x and 3,400x more energy efficient

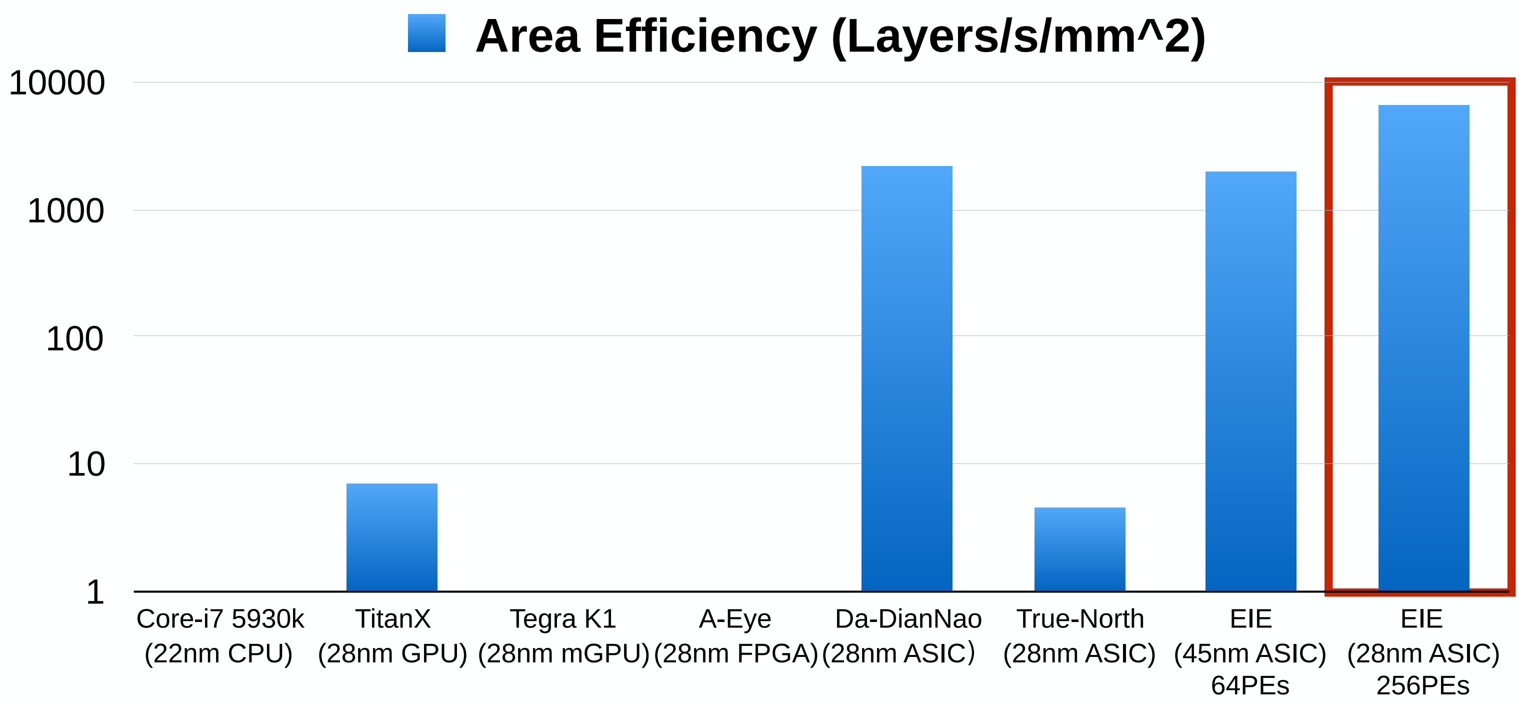
Baseline:

- Intel Core i7 5930K: reported by pcm-power utility
- NVIDIA GeForce GTX Titan X: reported by nvidia-smi utility
- NVIDIA Tegra K1: measured with power-meter, 60% AP+DRAM power

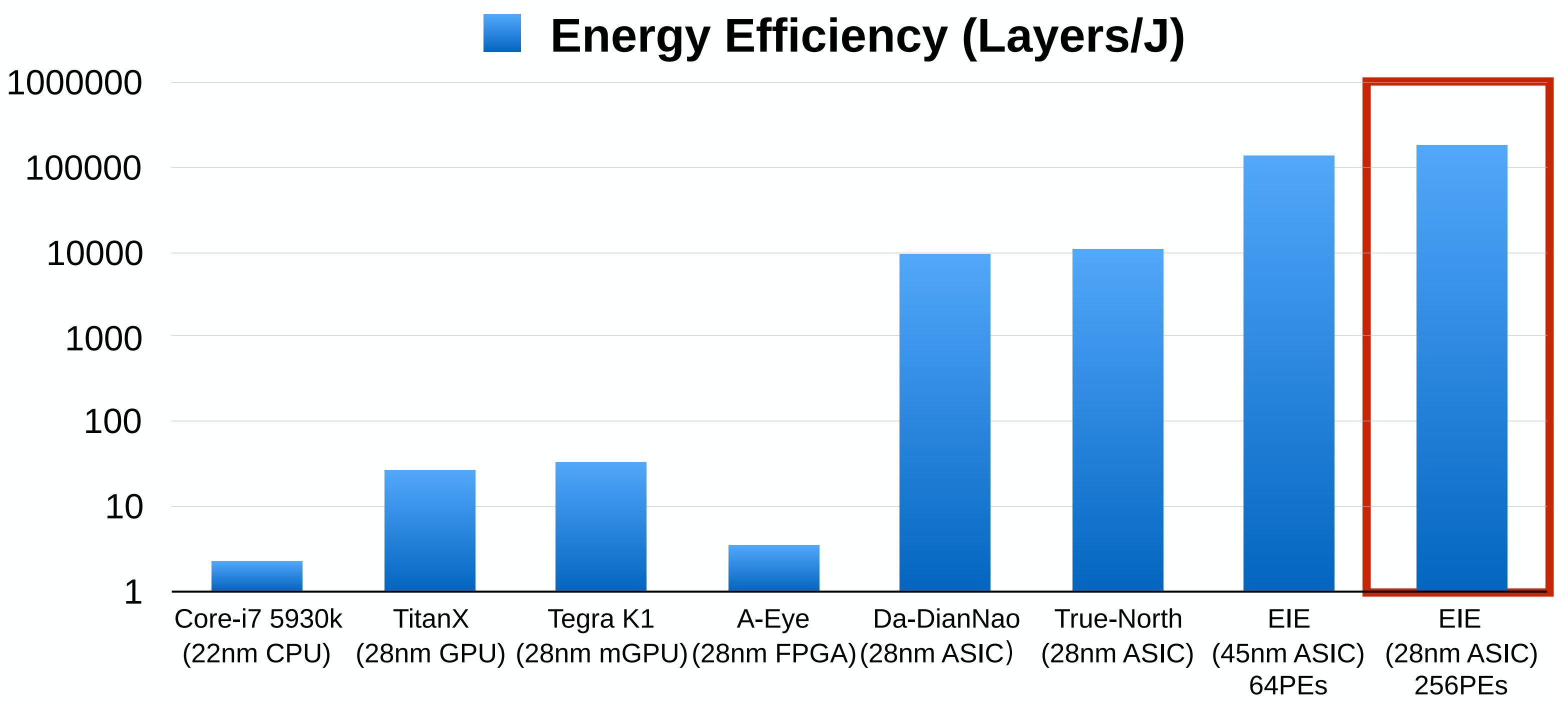
Comparison: Throughput



Comparison: Area Efficiency



Comparison: Energy Efficiency



Where are the savings from?

- <u>Four factors for energy saving:</u>

- **10× *static* weight sparsity;**
  less work to do; less bricks to carry.

- **3× *dynamic* activation sparsity;**
  carry only good bricks; ignore broken bricks.

- **Weight sharing with only 4-bits per weight;**
  lighter bricks to carry.

- **DRAM => SRAM, no need to go off-chip;**
  carry bricks from San Francisco to Seoul => Incheon to Seoul.

Conclusion

- EIE: first accelerator for compressed, sparse neural network.
- Compression => Acceleration, no loss accuracy.
- Distributed storage/computation to parallelize/load balance across PEs.
- 13x faster and 3,400x more energy efficient than GPU. 2.9x faster and 19x more energy efficient than past ASIC.

Beyond EIE: a Multi-Dimension Sparse Recipe for Deep Learning

Faster Speed: EIE accelerator

sparsity

Smaller Size: Deep Compression, SqueezeNet++

Higher Accuracy: DSD regularization

[1]. **Han** et al. “Learning both Weights and Connections for Efficient Neural Networks”, NIPS 2015

[2]. **Han** et al. “Deep Compression: Compressing Deep Neural Networks with Pruning, Trained Quantization and Huffman Coding”, Deep Learning Symposium 2015, ICLR 2016 (best paper award)

[3]. **Han** et al. “EIE: Efficient Inference Engine on Compressed Deep Neural Network”, ISCA 2016

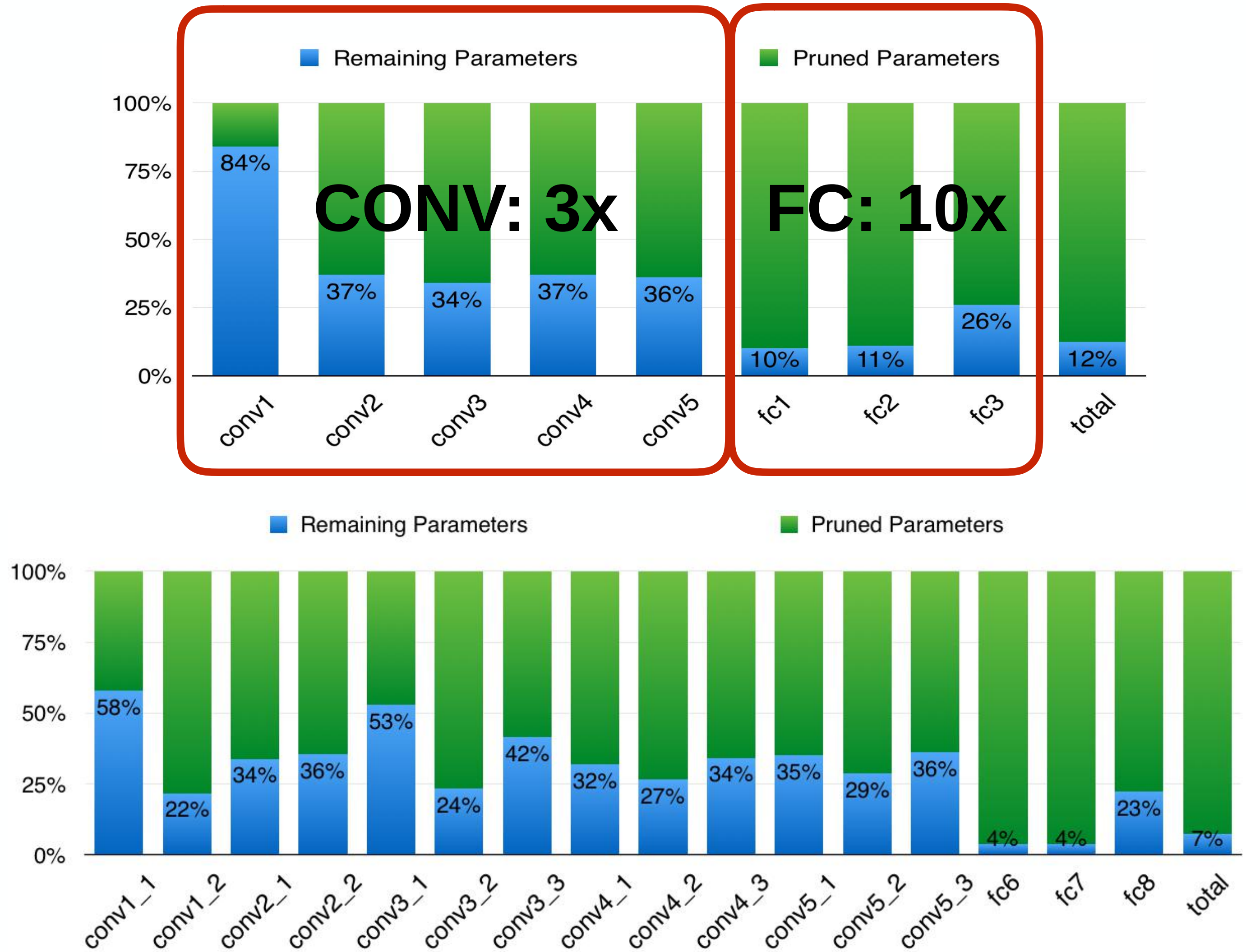
[4]. **Han** et al. “DSD: Regularizing Deep Neural Networks with Dense-Sparse-Dense Training Flow”, arXiv 2016

[5]. Iandola, **Han**,et al. “SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5MB model size”, arXiv’16

[6]. Yao, **Han**, et.al, “Hardware-friendly convolutional neural network with even-number filter size”, ICLR workshop 2016

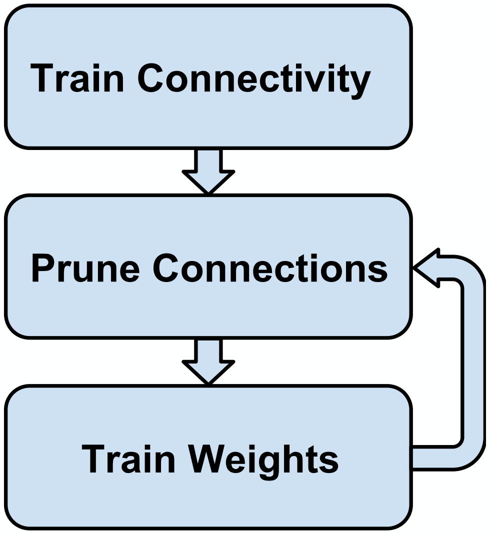
Backup Slides

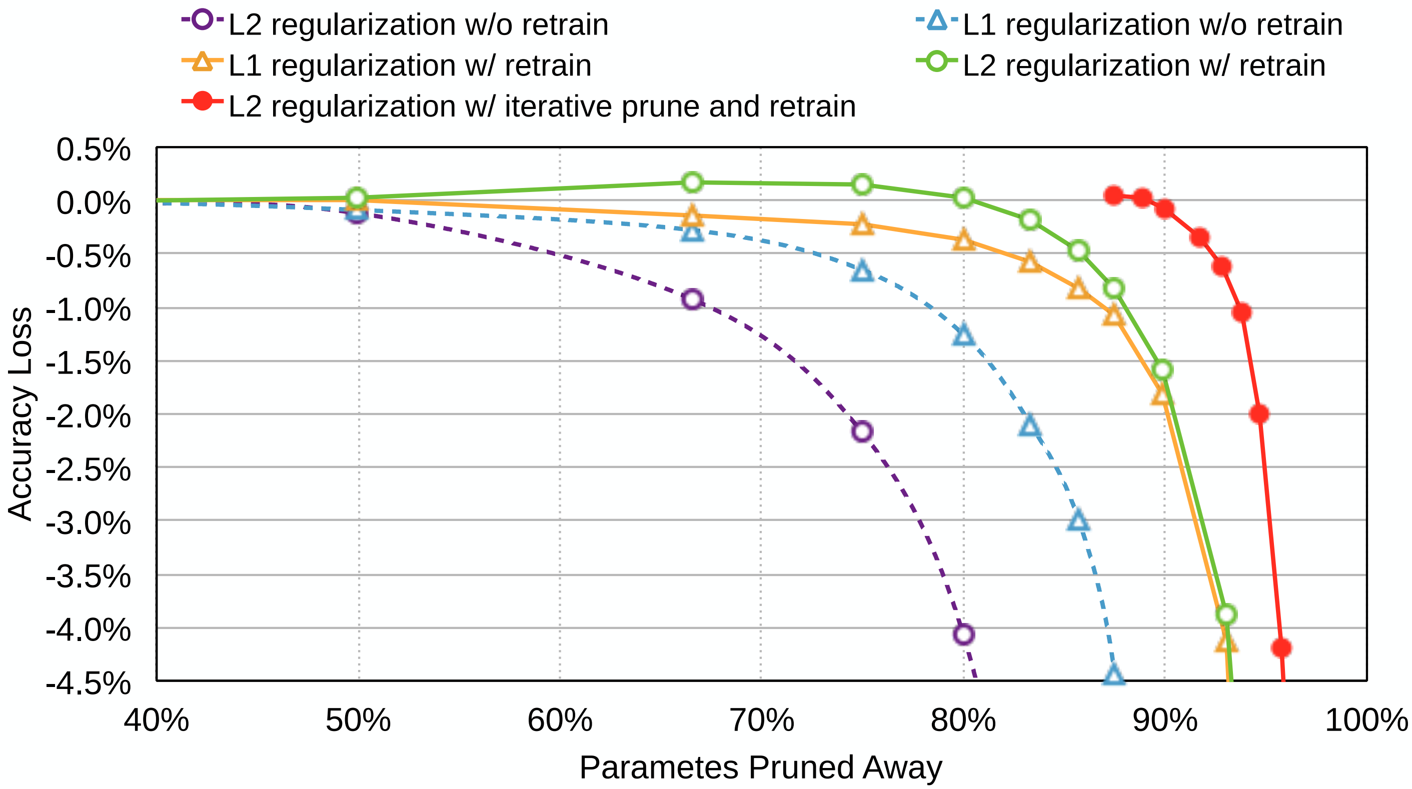
Sparsity: Pruning AlexNet & VGGNet



Han et al. "Learning both Weights and Connections for Efficient Neural Networks", NIPS 2015

Retrain to Fully Recover Accuracy





Han et al. “Learning both Weights and Connections for Efficient Neural Networks”, NIPS 2015

Weight Sharing: Accuracy with # Bits

#CONV bits / #FC bits	Top-1 Error	Top-5 Error	Top-1 Error Increase	Top-5 Error Increase
32bits / 32bits	42.78%	19.73%	-	-
8 bits / 5 bits	42.78%	19.70%	0.00%	-0.03%
8 bits / 4 bits	42.79%	19.73%	0.01%	0.00%
4 bits / 2 bits	44.77%	22.33%	1.99%	2.60%

Han et al. “Deep Compression: Compressing Deep Neural Networks with Pruning, Trained Quantization and Huffman Coding” ICLR 2016

Deep Compression Result on Major Convnets

Network	Top-1 Error	Top-5 Error	Parameters	Compress Rate
LeNet-300-100 Ref	1.64%	-	1070 KB	
LeNet-300-100 Compressed	1.58%	-	**27 KB**	**40**×
LeNet-5 Ref	0.80%	-	1720 KB	
LeNet-5 Compressed	0.74%	-	**44 KB**	**39**×
AlexNet Ref	42.78%	19.73%	240 MB	
AlexNet Compressed	42.78%	19.70%	**6.9 MB**	**35**×
VGG-16 Ref	31.50%	11.32%	552 MB	
VGG-16 Compressed	31.17%	10.91%	**11.3 MB**	**49**×
SqueezeNet Ref	42.5%	19.7%	4.8 MB	
SqueezeNet Compressed	42.5%	19.7%	**0.47MB**	**10**×
GoogLeNet Ref	31.30%	11.10%	28 MB	
GoogLeNet Compressed	31.26%	11.08%	**2.8 MB**	**10**×

- SqueezeNet and GoogleNet: just Pruning and Quantization gives 10x compression.
- Inception Model is really efficient for classification.
- But it can still achieve an order of magnitude smaller with Deep Compression.
- **Fits in SRAM cache.**

Han et al. “Deep Compression: Compressing Deep Neural Networks with Pruning, Trained Quantization and Huffman Coding” ICLR 2016

Pruning NeuralTalk and LSTM

- **Original**: a basketball player in a white uniform is playing with a ball
- **Pruned 90%**: a basketball player in a white uniform is playing with a basketball

- **Original** : a brown dog is running through a grassy field
- **Pruned 90%**: a brown dog is running through a grassy area

- **Original** : a man is riding a surfboard on a wave
- **Pruned 90%**: a man in a wetsuit is riding a wave on a beach

- **Original** : a soccer player in red is running in the field
- **Pruned <u>95%</u>**: a man in a red shirt and black and white black shirt is running through a field

Han et al. "Learning both Weights and Connections for Efficient Neural Networks", NIPS 2015 poster

With Sparsity Constraint, DSD Training Improves Accuracy (Baseline: NeuralTalk)

Baseline: a boy is swimming in a pool.
Sparse: a small black dog is jumping into a pool.
DSD: a black and white dog is swimming in a pool.

Baseline: a group of people are standing in front of a building.
Sparse: a group of people are standing in front of a building.
DSD: a group of people are walking in a park.

Baseline: two girls in bathing suits are playing in the water.
Sparse: two children are playing in the sand.
DSD: two children are playing in the sand.

Baseline: a man in a red shirt and jeans is riding a bicycle down a street.
Sparse: a man in a red shirt and a woman in a wheelchair.
DSD: a man and a woman are riding on a street.

Baseline: a group of people sit on a bench in front of a building.
Sparse: a group of people are standing in front of a building.
DSD: a group of people are standing in a fountain.

Baseline: a man in a black jacket and a black jacket is smiling.
Sparse: a man and a woman are standing in front of a mountain.
DSD: a man in a black jacket is standing next to a man in a black shirt.

Baseline: a group of football players in red uniforms.
Sparse: a group of football players in a field.
DSD: a group of football players in red and white uniforms.

Baseline: a dog runs through the grass.
Sparse: a dog runs through the grass.
DSD: a white and brown dog is running through the grass.

Han et al. “DSD: Regularizing Deep Neural Networks with Dense-Sparse-Dense Training Flow”, arXiv 2016