Deep Compression and EIE: Efficient Inference Engine on Compressed Deep Neural Network

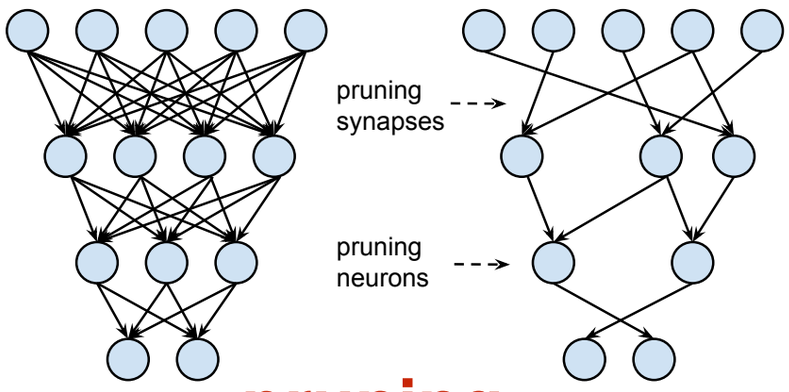
Song Han*, Xingyu Liu, Huizi Mao, Jing Pu, Ardavan Pedram,
Mark Horowitz, Bill Dally

Stanford University

Our Prior Work: Deep Compression

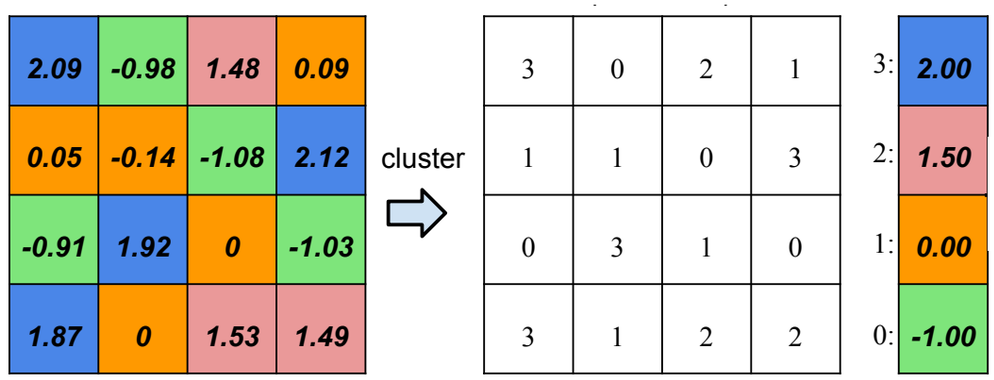
- **Memory reference is expensive.**
- **Small DNN models are critical.**

[1]. Han et al. NIPS 2015



pruning

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



weight sharing

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%

EIE: First Accelerator for Sparse DNN

- **Deep Compression solves the model size problem.**
- **But it creates another problem: irregular computation pattern.**
- **CPU/GPU are only good at dense linear algebra.**
- **So we create EIE that supports: static-sparse M, dynamic-sparse V, indirect indexing, weight sharing.**

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

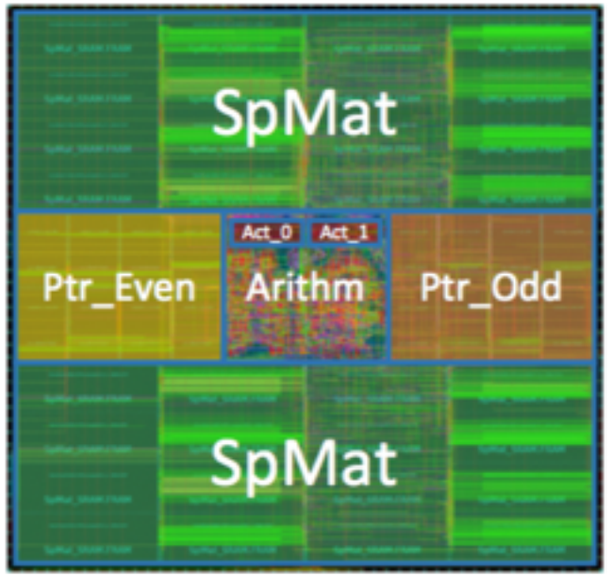
Fully fits in SRAM

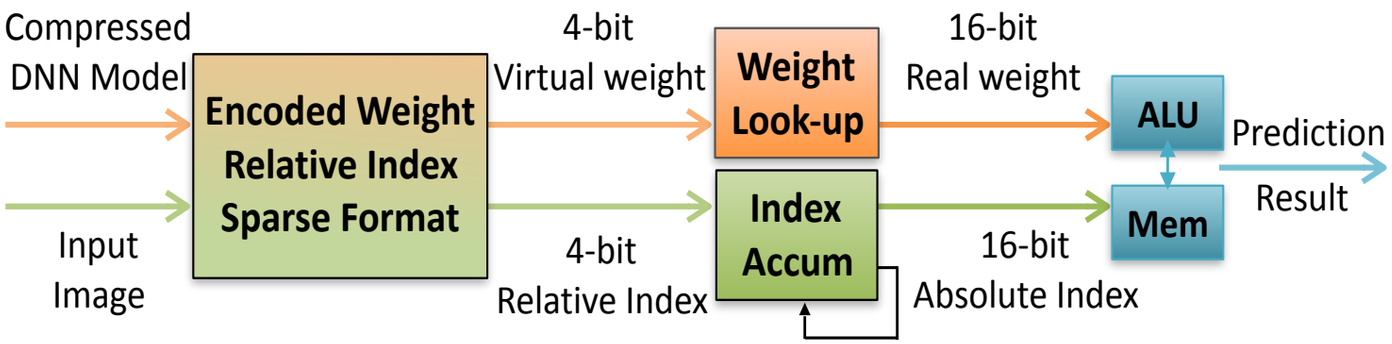
120x less energy than DRAM

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

Dally. NIPS tutorial 2015; Han et al. ISCA 2016

EIE: First Accelerator for Sparse DNN



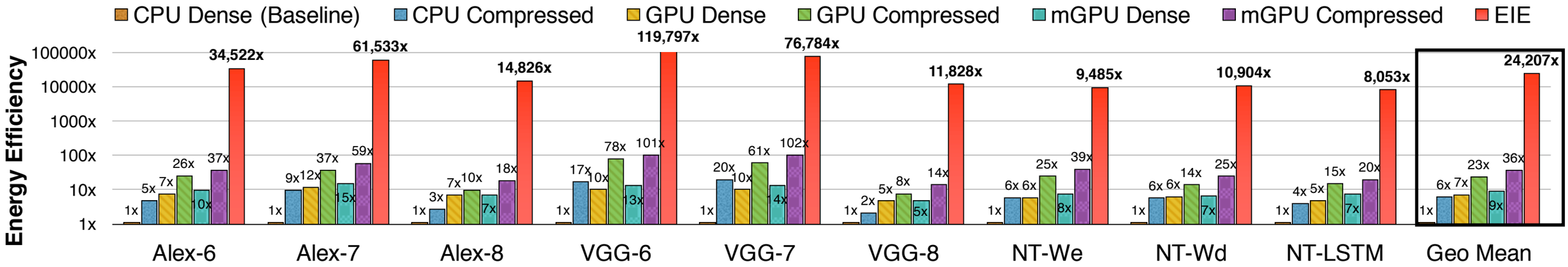


Technology	45 nm
# PEs	64
on-chip SRAM	8 MB
Max Model Size	84 Million
Static Sparsity	10x
Dynamic Sparsity	3x
Quantization	4-bit
ALU Width	16-bit
Area	40.8 mm^2
MxV Throughput	81,967 layers/s
Power	586 mW

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

Dally. NIPS tutorial 2015; Han et al. ISCA 2016

FC Layers: Speedup / Energy Efficiency



Compared to CPU and GPU:

189x and 13x faster

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

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

Faster Speed: EIE accelerator

[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