Designing Energy-Efficient Convolutional Neural Networks using Energy-Aware Pruning

Tien-Ju Yang, Yu-Hsin Chen, Vivienne Sze

Massachusetts Institute of Technology

http://eyeriss.mit.edu/energy.html

IEEE 2017 Conference on Computer Vision and Pattern Recognition

CVPR July 21-26 Honolulu 2017

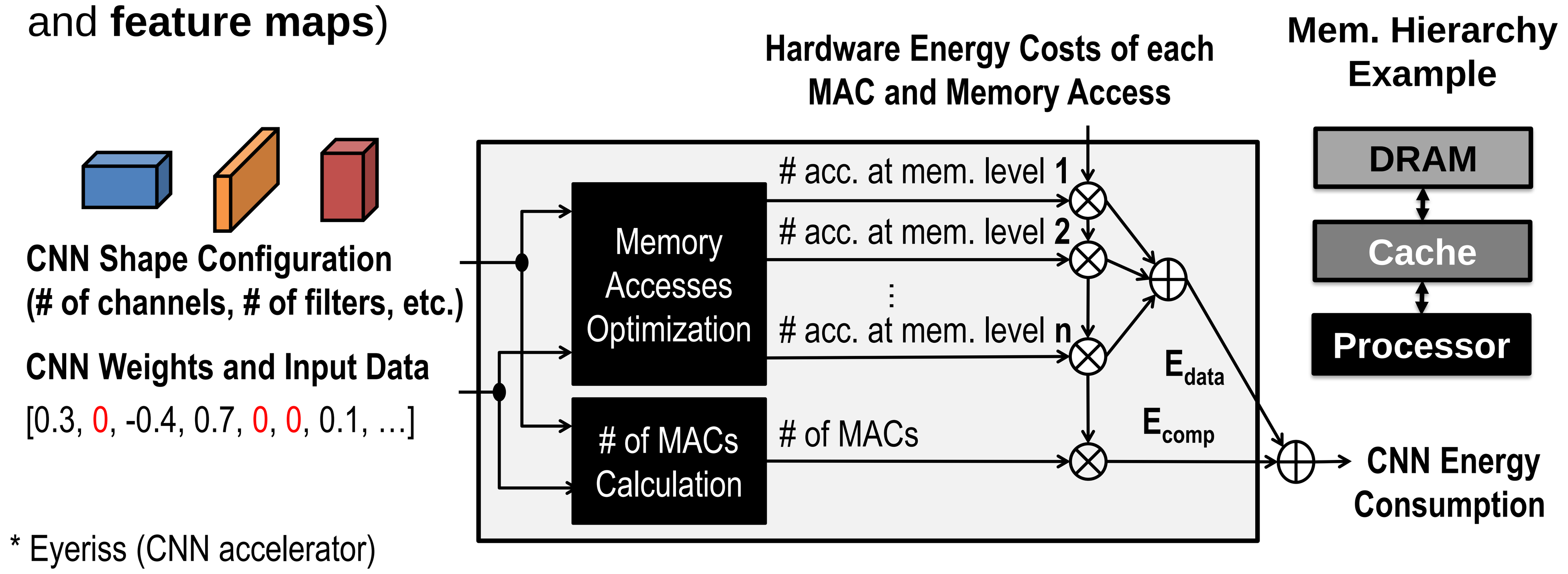
CNN – Accurate but High Energy Consumption

- High energy consumption hinders CNN deployment on battery-powered devices
- We propose an energy-aware pruning algorithm for CNNs that directly targets energy rather than number of weights
- We perform energy analysis on various CNNs and provide insights

Energy Estimation Methodology

* The tool is available on the website

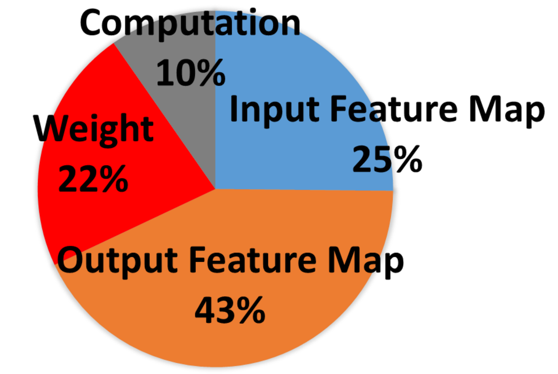
- Energy consumption: a combination of # of memory accesses and # of MACs (energy model from hardware measurements*)
- Energy model considers bit-width and sparsity, as well as all data types (weights and **feature maps**)



* Eyeriss (CNN accelerator)

Energy Consumption Analysis of CNNs

- Number of weights is not a good estimator of energy
- Example 1: CONV layers consume more energy than FC layers
- Example 2: deeper CNNs with fewer weights do not necessarily consume less energy than shallower CNNs with more weights



GoogLeNet Energy

Example 1

Layer	# Weights	Energy
CONV	4%	73%
FC	96%	27%

AlexNet

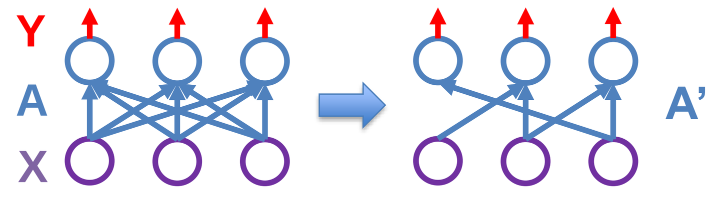
Example 2

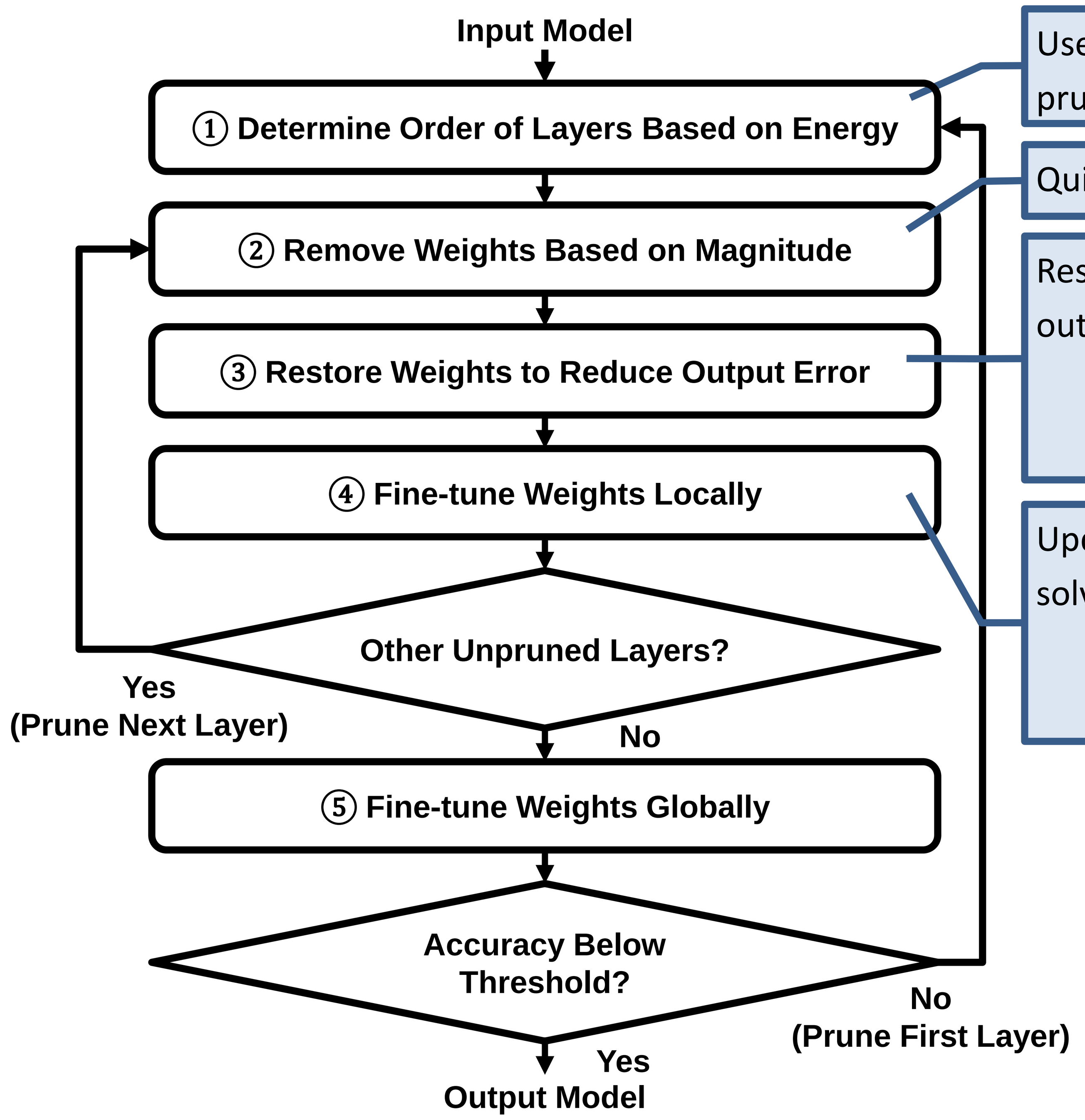
Network	# Weights	Energy
AlexNet	61M	4G
GoogLeNet	7M	8G

AlexNet vs. GoogLeNet

Energy-Aware Pruning Algorithm

- Reduce the energy consumption by pruning a network
- Focus on minimizing the output error instead of the filter error





Use the estimated energy to guide the pruning algorithm

Quickly remove weights

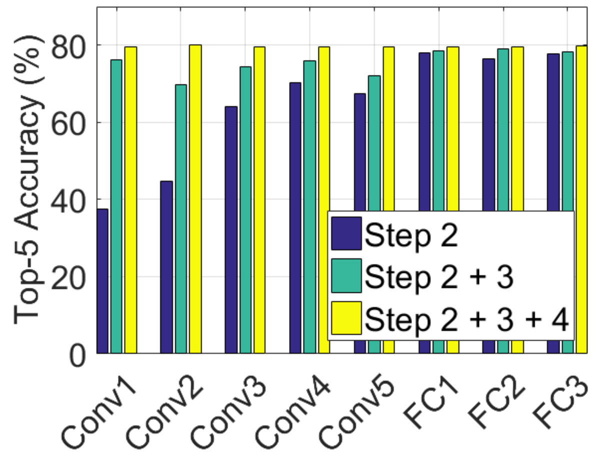
Restore weights to minimize the output feature map error:

$$A' = \arg\min_{A} \|Y - XA\|_p^p, \quad \text{subject to } \|A\|_0 \leq q$$

Update the remaining weights (A_S') by solving the least-square problem:

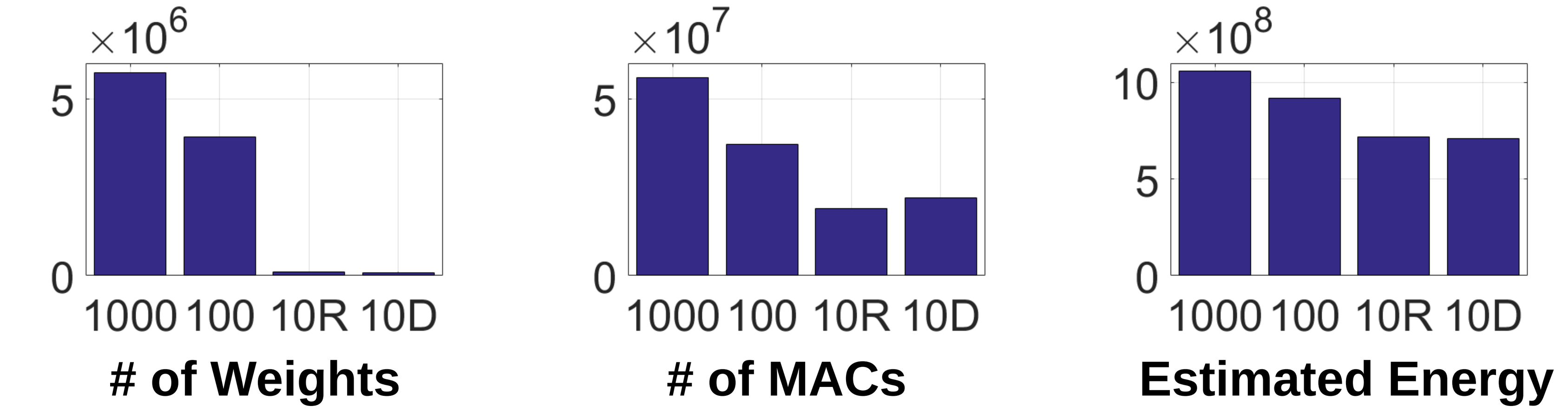
$$A_S' = \arg\min_{A_S} \|Y - X_S A_S\|_2^2$$

$$A_{S^c}' = 0$$



Reducing Number of Target Classes

- Key observation: by reducing the number of target classes on AlexNet, the model size is greatly reduced but the energy reduction is limited



Pruning Results for AlexNet and GoogLeNet

- **3.7x** and **1.6x** energy reduction
- **10.6x** and **3.0x** number of non-zero weights reduction
- **6.6x** and **3.4x** number of non-skipped MACs reduction

Model	Top-5 Accuracy	# of Non-zero Weights (×10⁶)		# of Non-skipped MACs (×10⁸)		Normalized Energy (×10⁹)	
AlexNet (Original)	80.43%	60.95	(100%)	3.71	(100%)	3.97	(100%)
AlexNet (DC)	80.37%	6.79	(11%)	1.79	(48%)	1.85	(47%)
AlexNet (This Work)	79.56%	5.73	(9%)	0.56	(15%)	1.06	(27%)
GoogLeNet (Original)	88.26%	6.99	(100%)	7.41	(100%)	7.63	(100%)
GoogLeNet (This Work)	87.28%	2.37	(34%)	2.16	(29%)	4.76	(62%)

- DC = S. Han et al., "Deep Compression: Compressing Deep Neural Networks with Pruning, Trained Quantization and Huffman Coding," in ICLR, 2016
- Energy-aware pruned models available for download from the website

Network Comparison (Energy vs. Accuracy)

- Our pruned networks achieve better accuracy-energy trade-off
- Feature maps need to be factored in when estimating energy