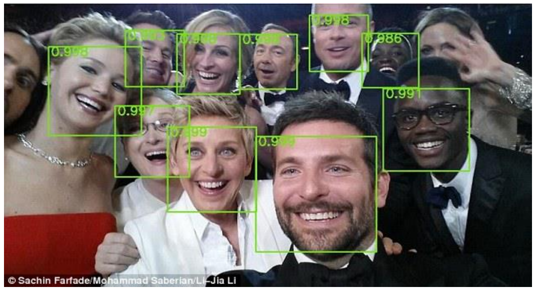
How to Estimate the Energy Consumption of Deep Neural Networks

Tien-Ju Yang, Yu-Hsin Chen,
Joel Emer, Vivienne Sze
MIT

Problem of DNNs



Recognition

Smart Drone

AI

Computation

DNN
15k – 300k OP/Px

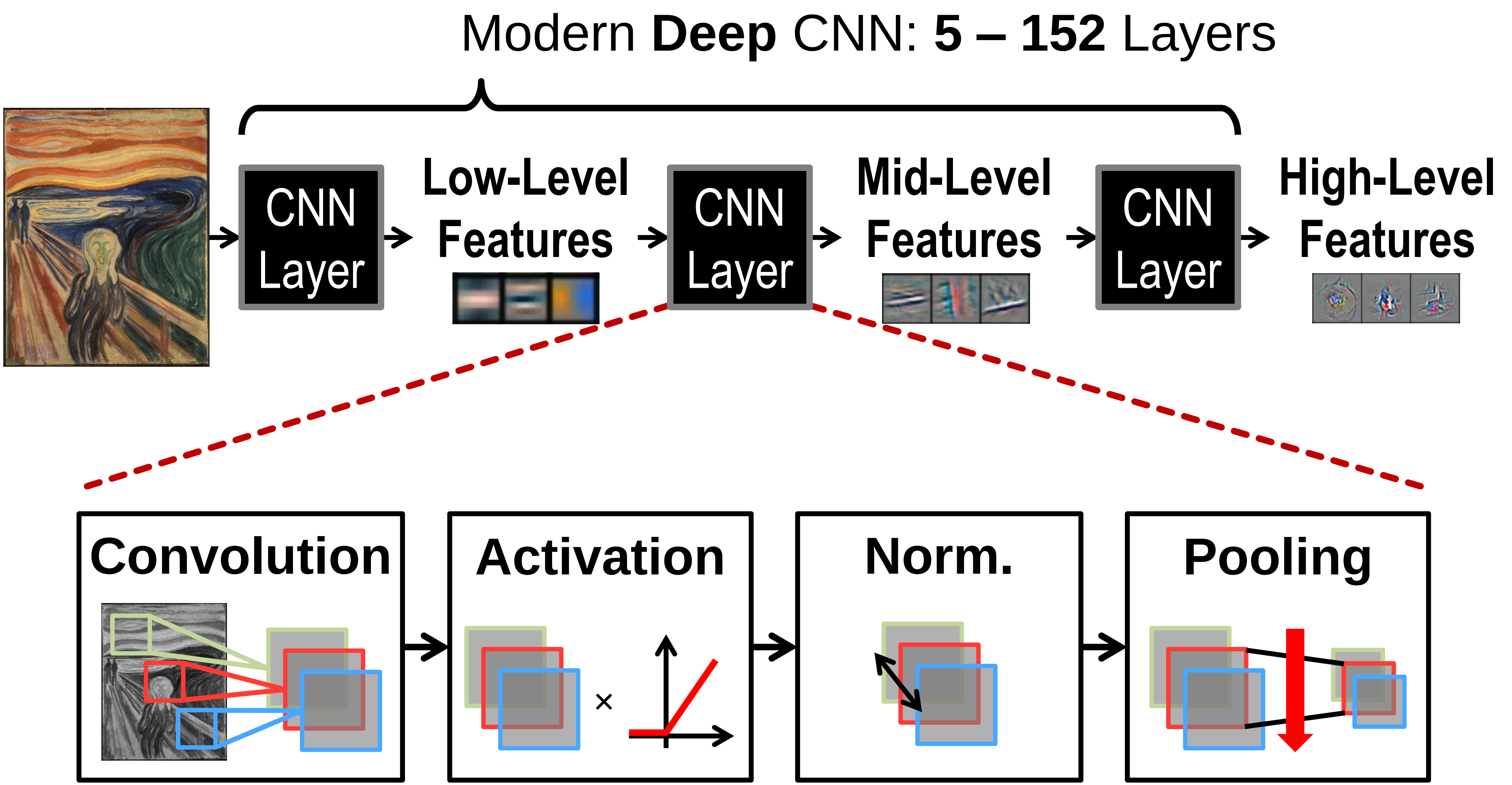
DPM
0.1k – 0.5k OP/Px

Accuracy

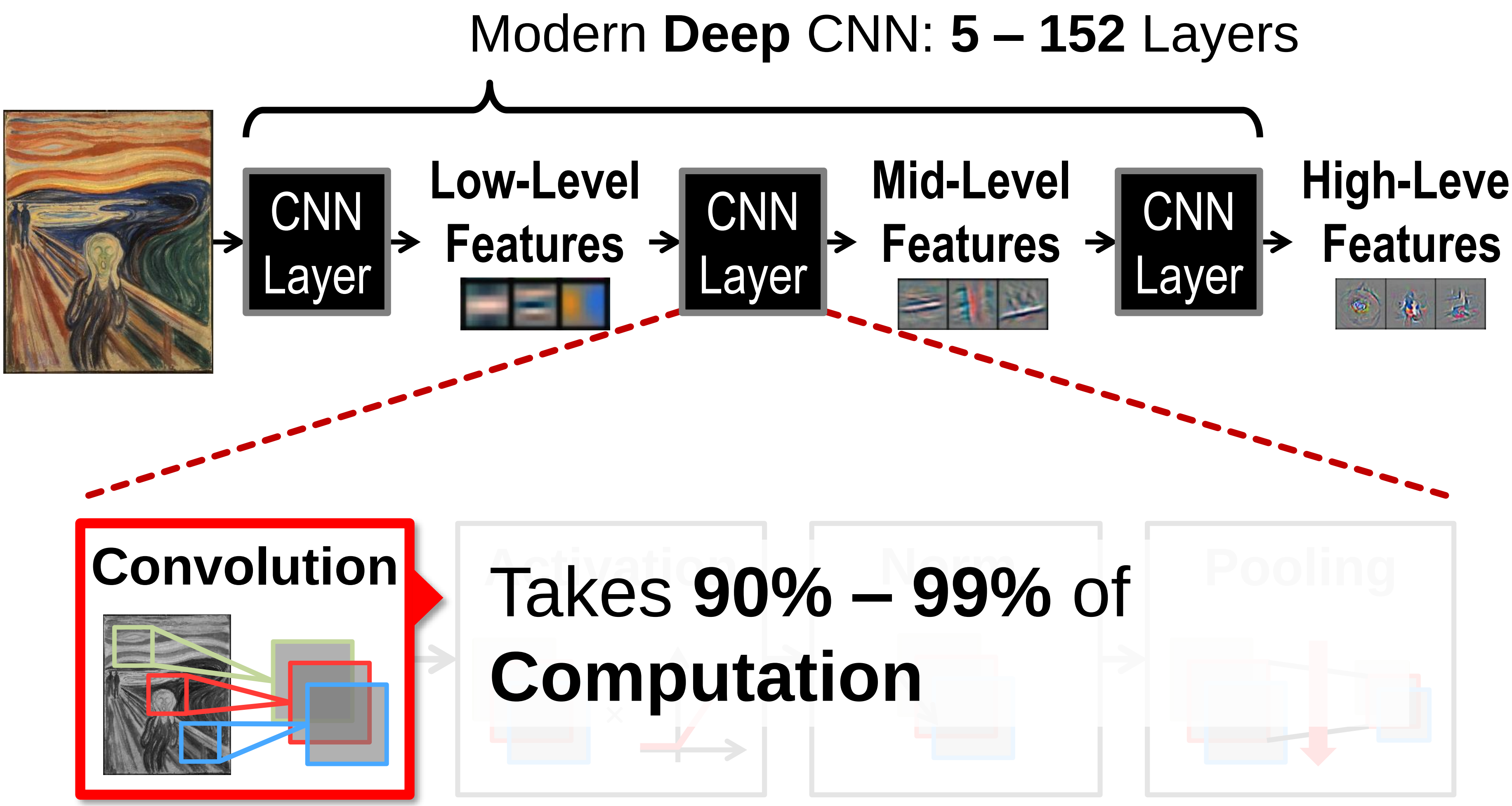
Energy Estimation helps

- **Understand** the design trade-off
- **Guide** the DNN design
- **Enable** DNN mobile applications

Deep Convolutional NN Explanation



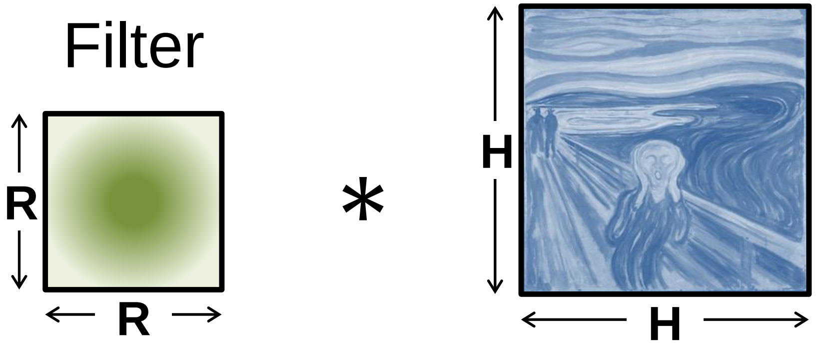
Deep Convolutional NN Explanation



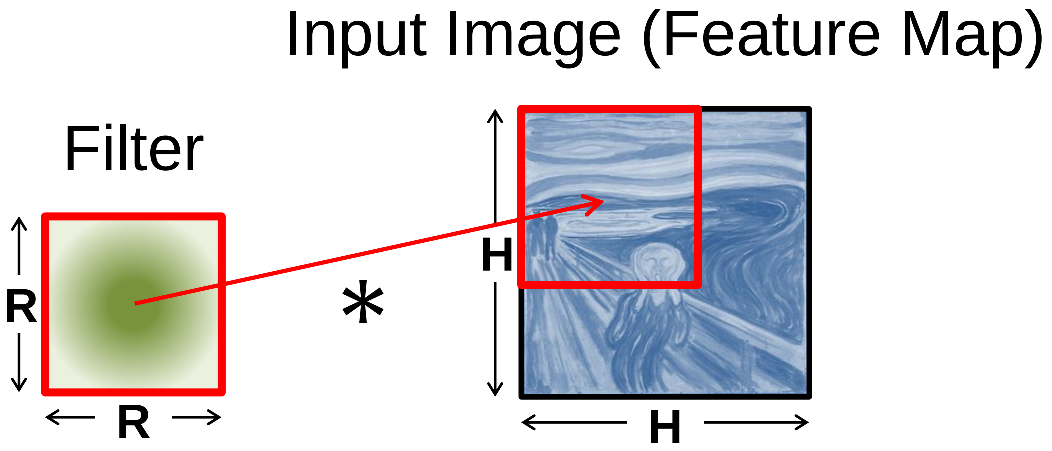
Convolution

Input Image (Feature Map)

Filter

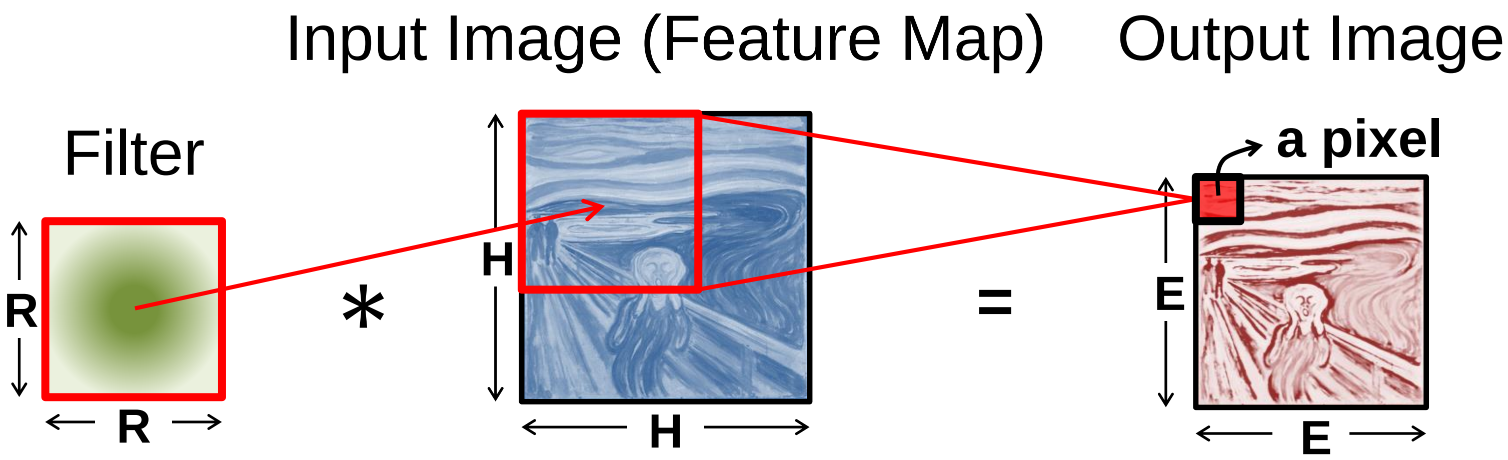


Convolution



Element-wise Multiplication

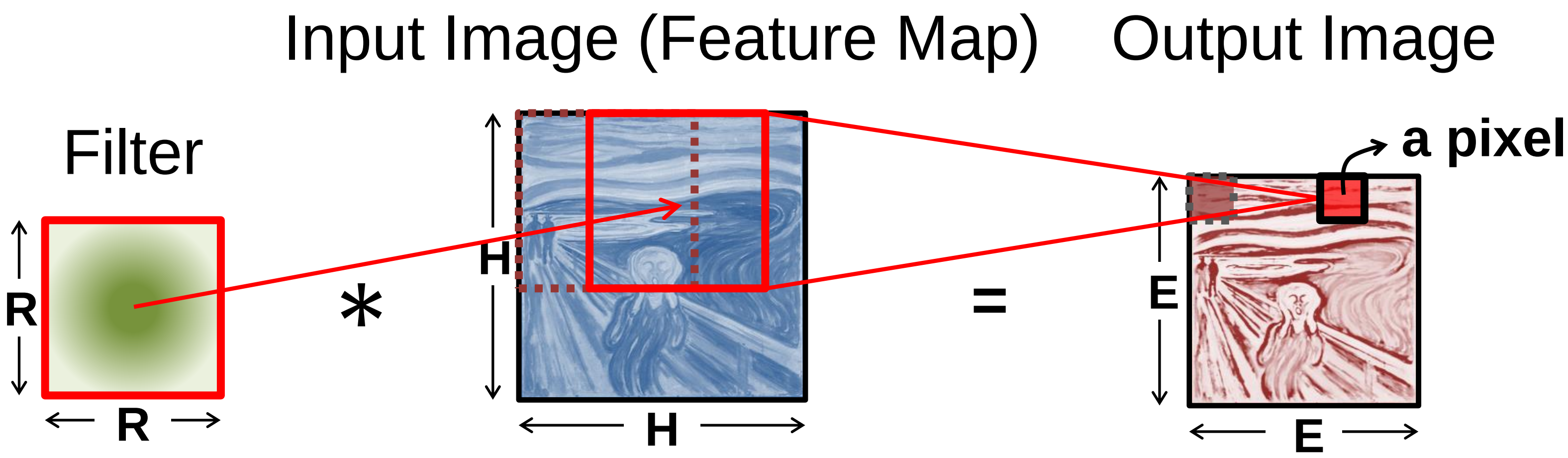
Convolution



Element-wise Multiplication

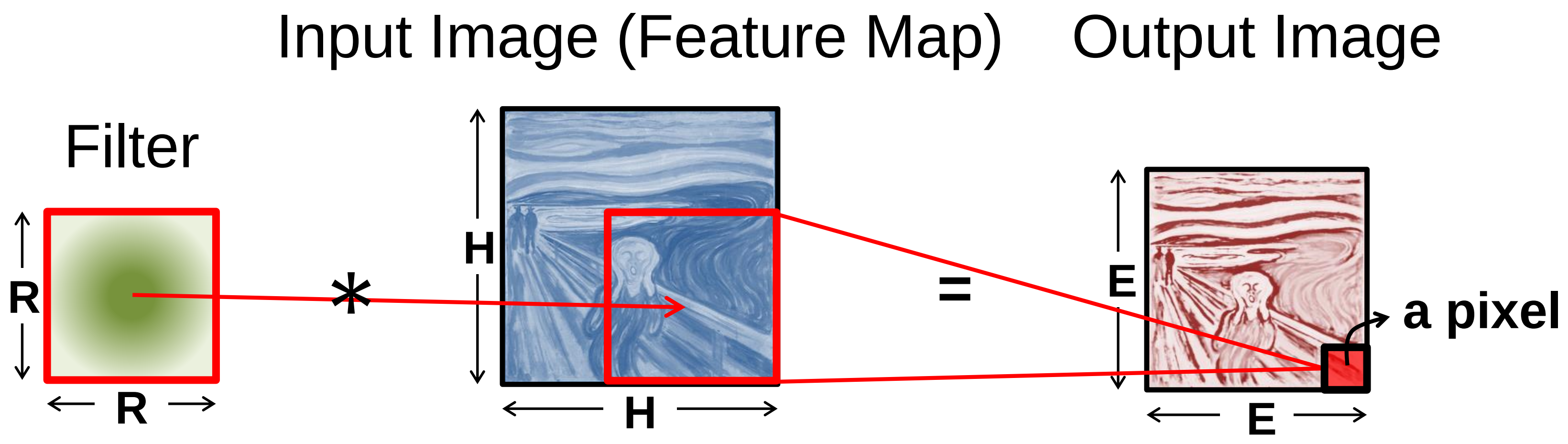
Partial Sum (psum) **Accumulation**

Convolution



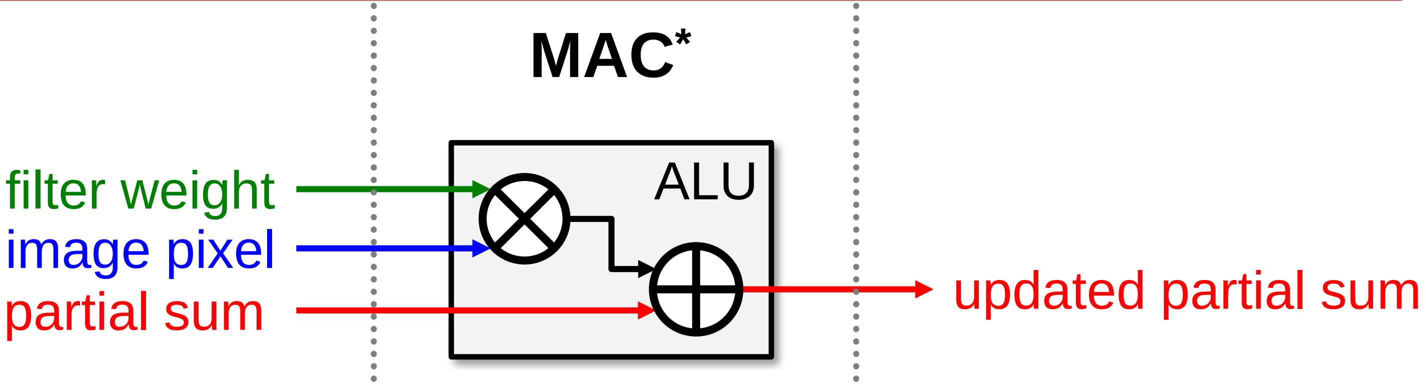
Sliding Window Processing

Convolution



Sliding Window Processing

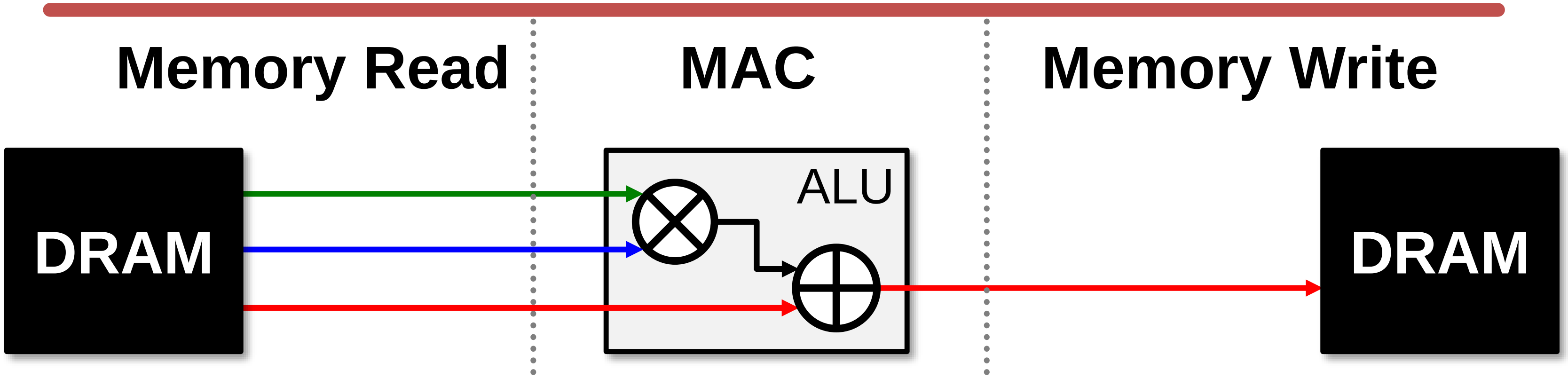
Why Not Use # of Weights or MACs?



* multiplication-and-accumulation

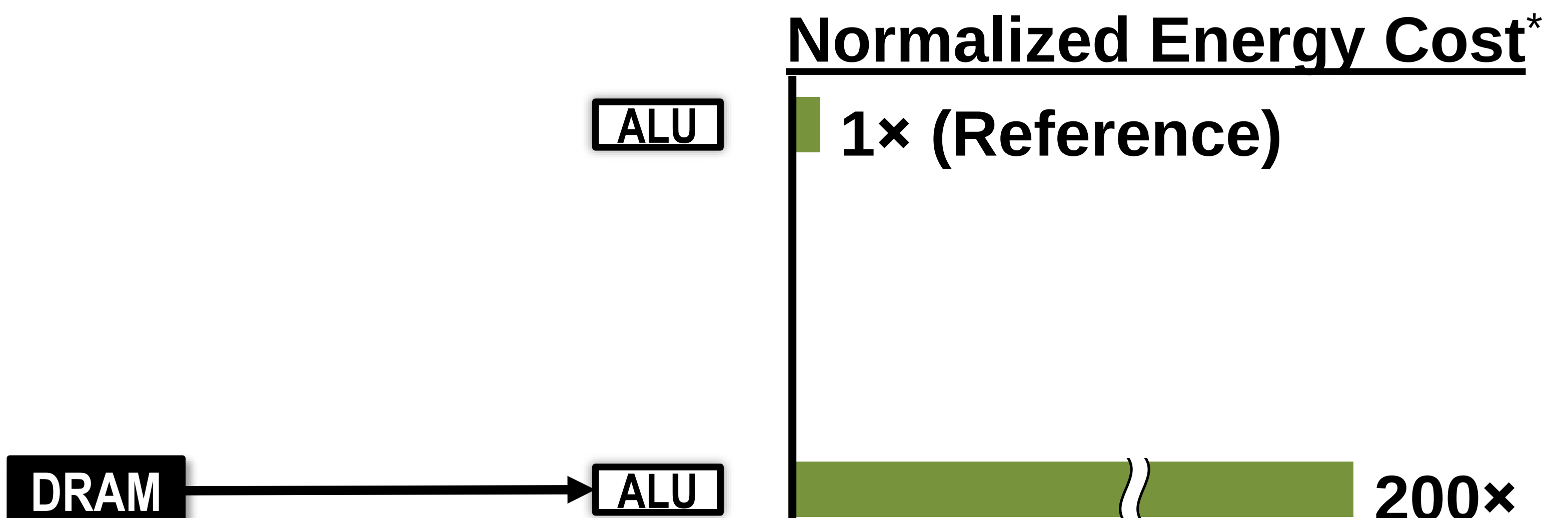
Reason 1:
Reason 2:

Why Not Use # of Weights or MACs?



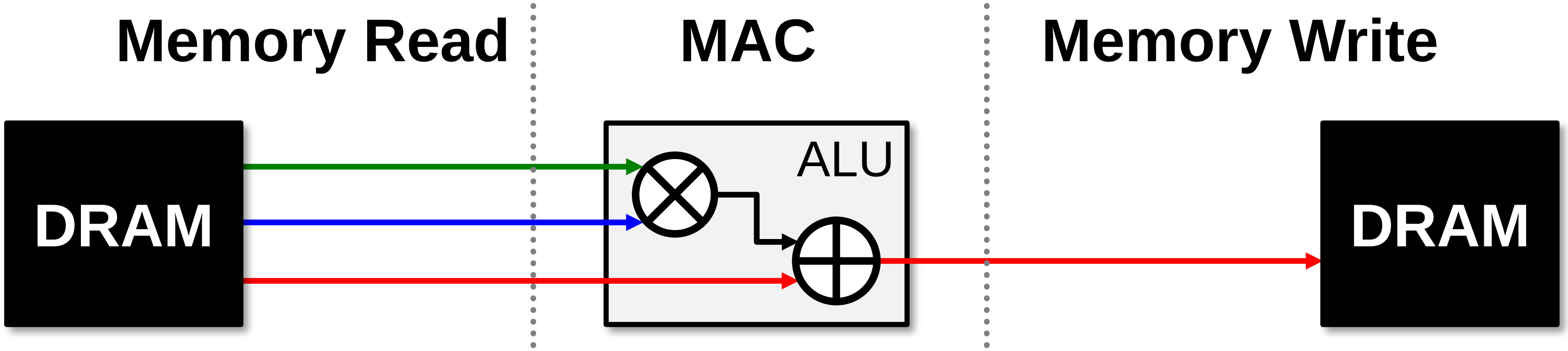
Reason 1:

Reason 2:



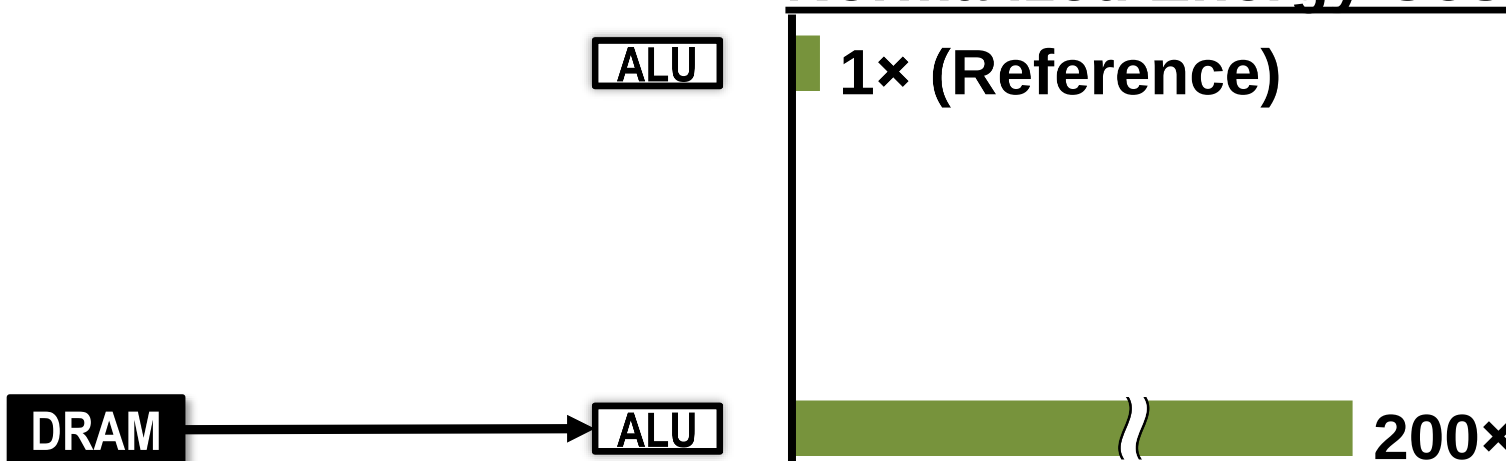
* measured from a commercial 65nm process

Why Not Use # of Weights or MACs?



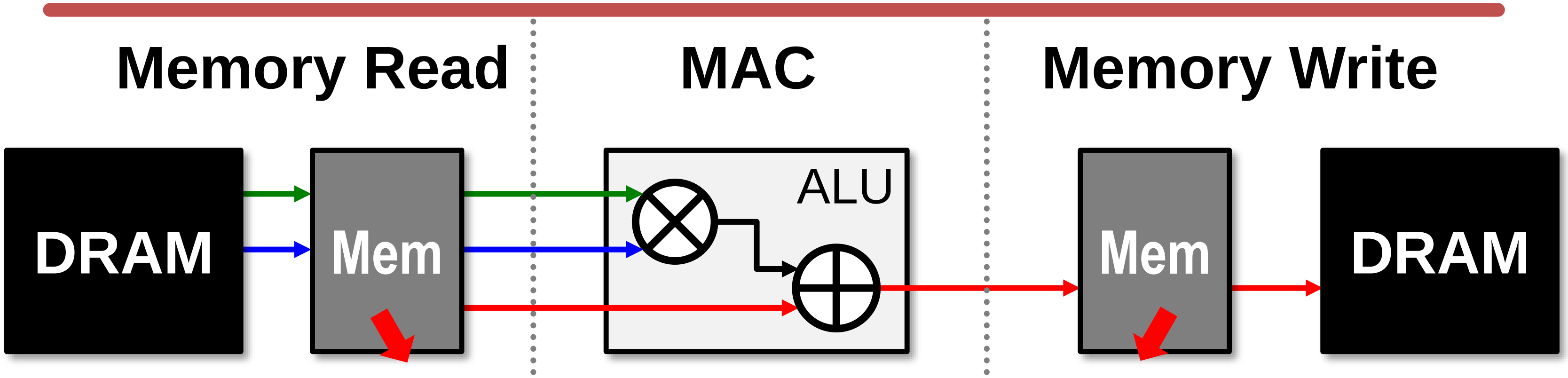
Reason 1: computation is cheap but **data movement** is expensive
Reason 2:

Normalized Energy Cost*



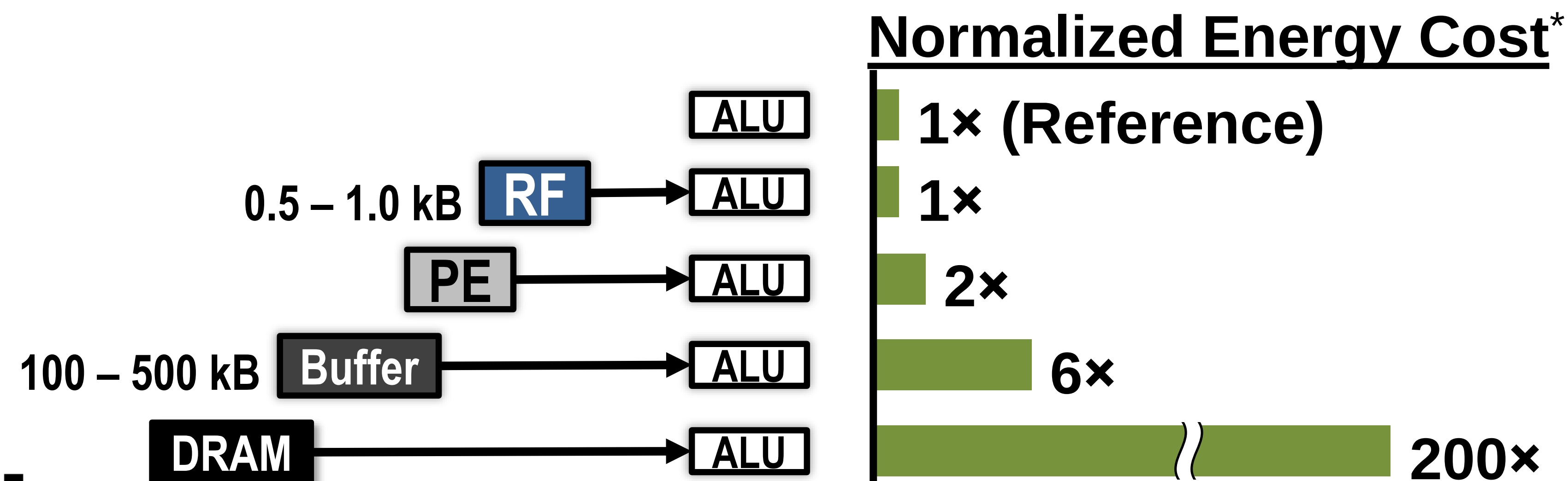
* measured from a commercial 65nm process

Why Not Use # of Weights or MACs?



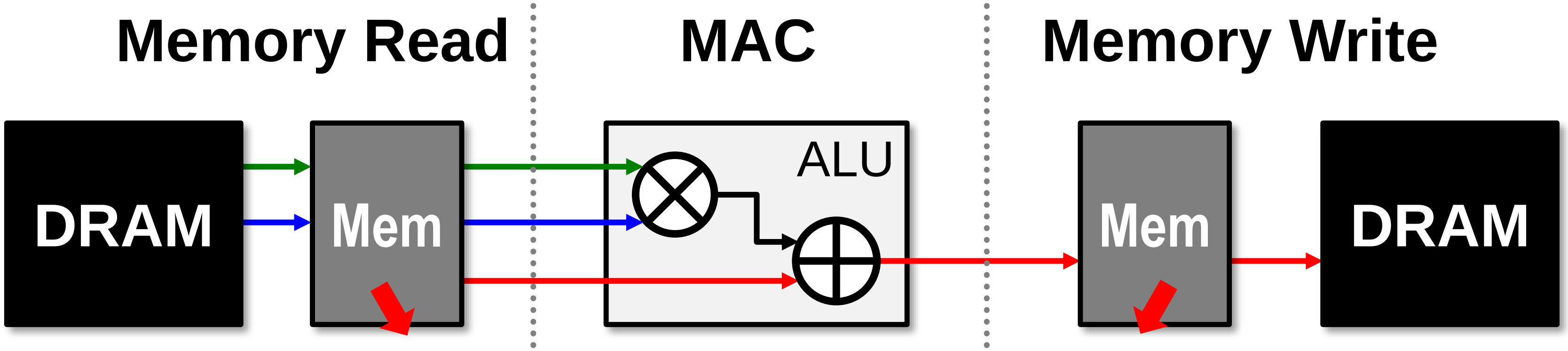
Extra levels of local memory hierarchy

Reason 1: computation is cheap but **data movement** is expensive
Reason 2:



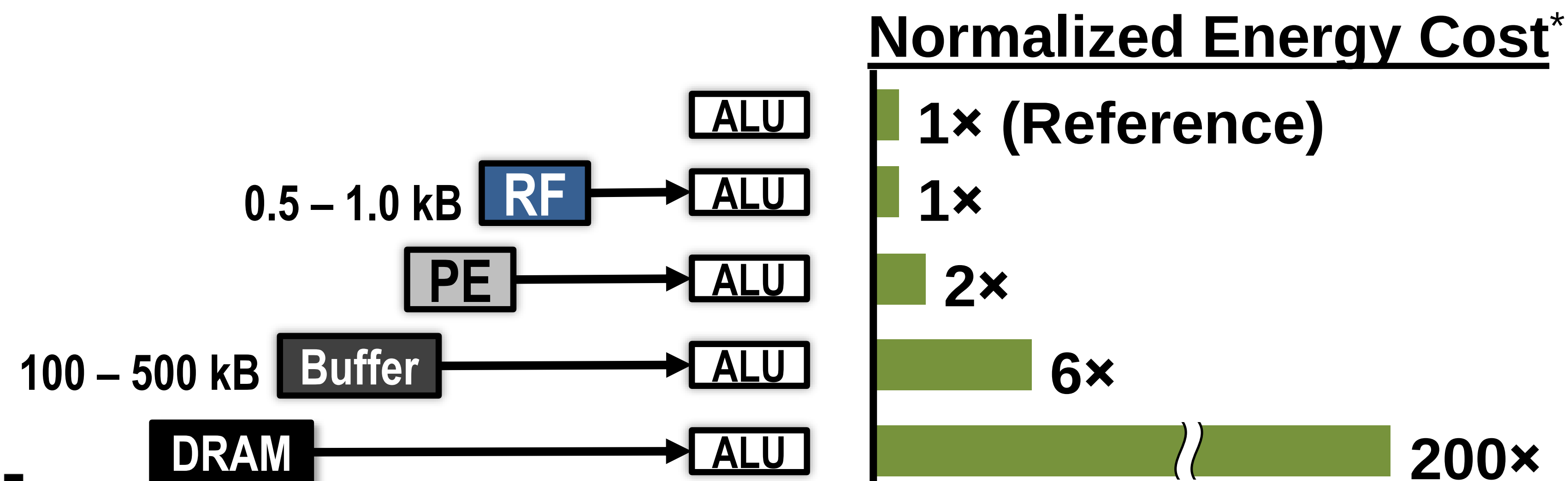
* measured from a commercial 65nm process

Why Not Use # of Weights or MACs?



Extra levels of local memory hierarchy

Reason 1: computation is cheap but **data movement** is expensive
Reason 2: where data come from/go to is important for energy

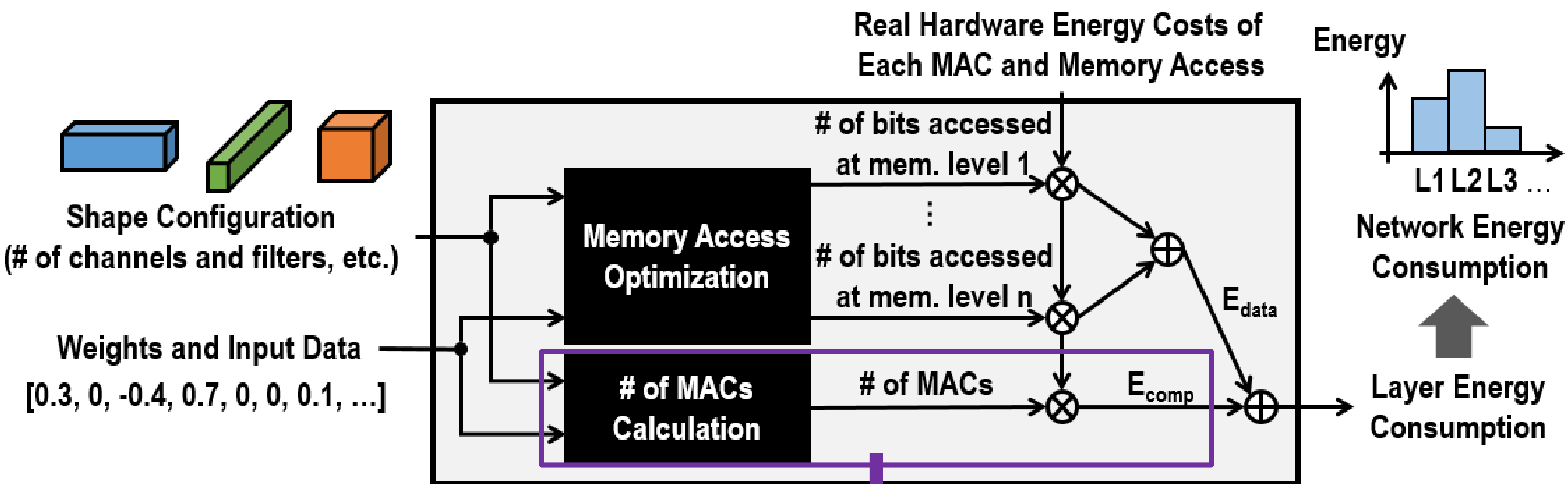


* measured from a commercial 65nm process

Energy Estimation Methodology

Energy Estimation Methodology

- Estimate the energy consumption of each layer separately
- For each layer, $E_{layer} = \boxed{E_{comp}} + E_{data}$

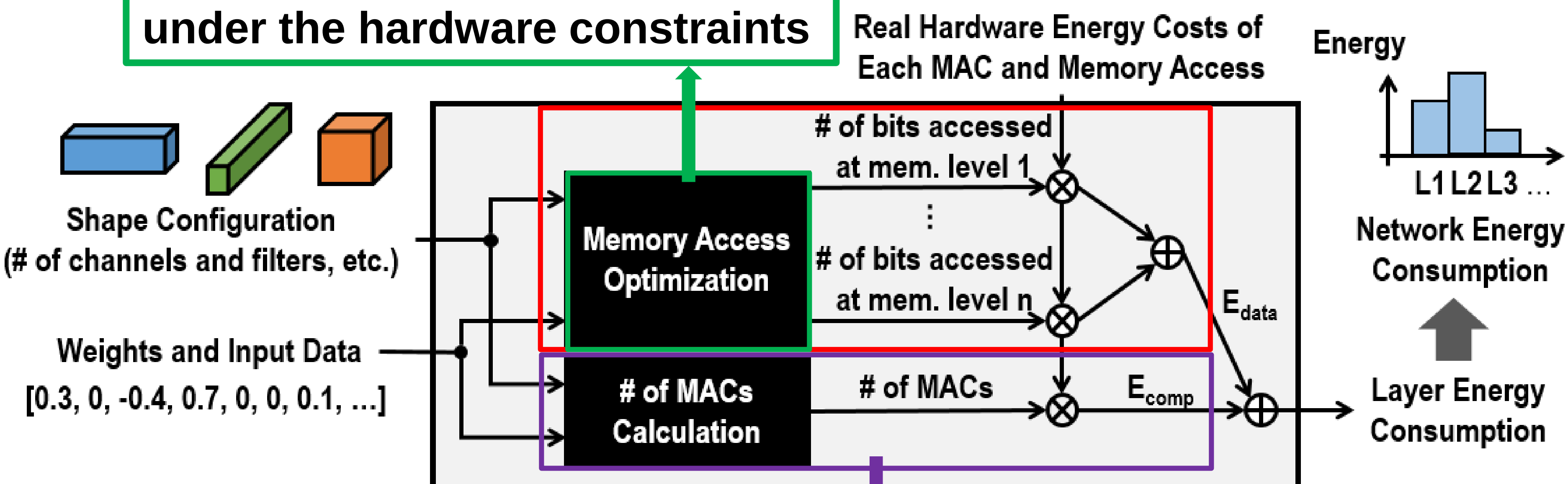


Computation energy only depends on the # of MACs

Energy Estimation Methodology

- Estimate the energy consumption of each layer separately
- For each layer, $E_{layer} = \boxed{E_{comp}} + \boxed{E_{data}}$

Minimize energy consumption under the hardware constraints

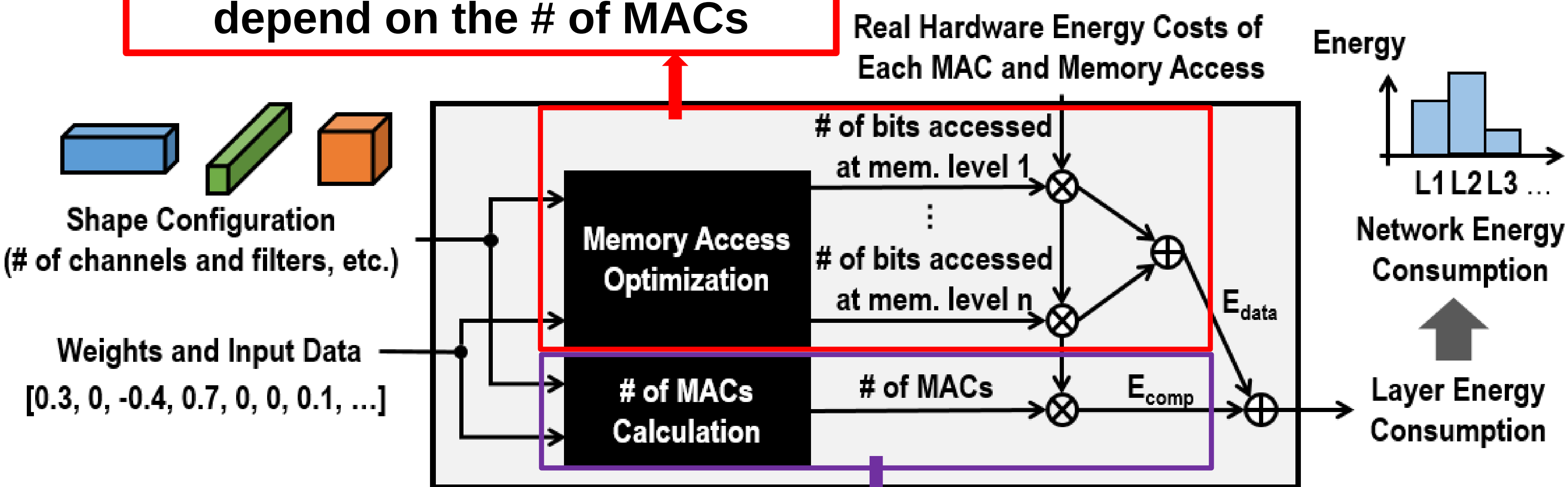


Computation energy only depends on the # of MACs

Energy Estimation Methodology

- Estimate the energy consumption of each layer separately
- For each layer, $E_{layer} = \boxed{E_{comp}} + \boxed{E_{data}}$

Data energy does NOT only depend on the # of MACs

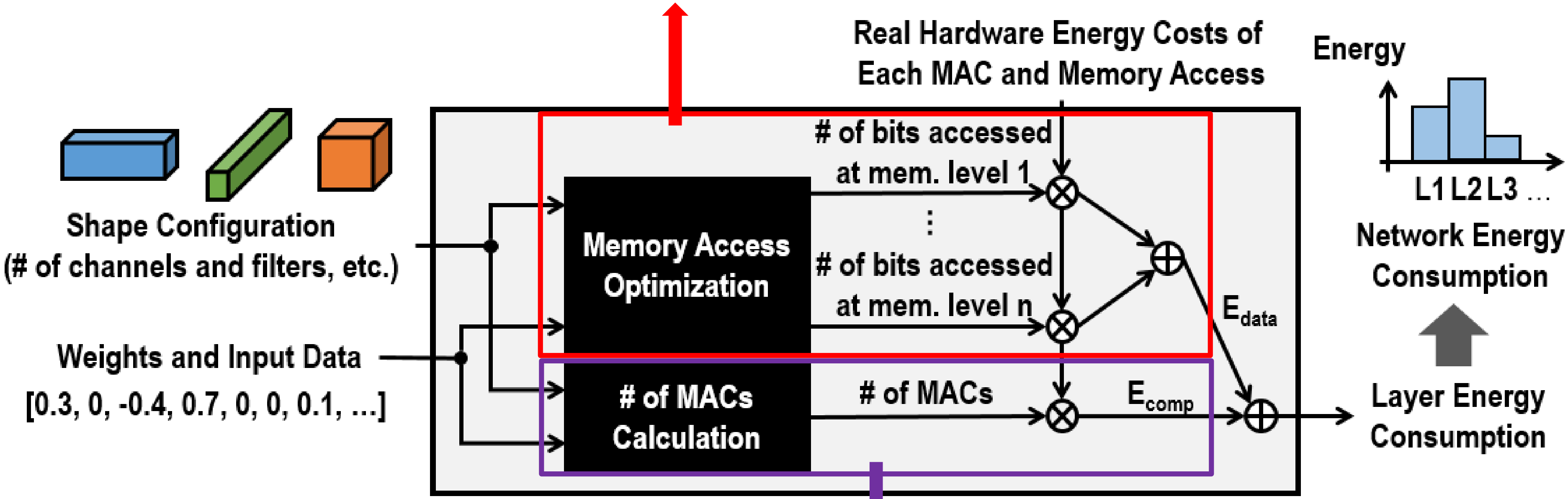


Computation energy only depends on the # of MACs

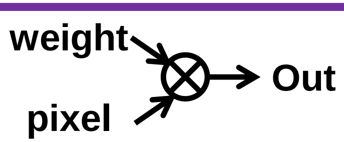
Factor in Bitwidth

Data energy:

- **Consider bitwidths in the optimization**
- **Scale # of bits linearly with the bitwidth**

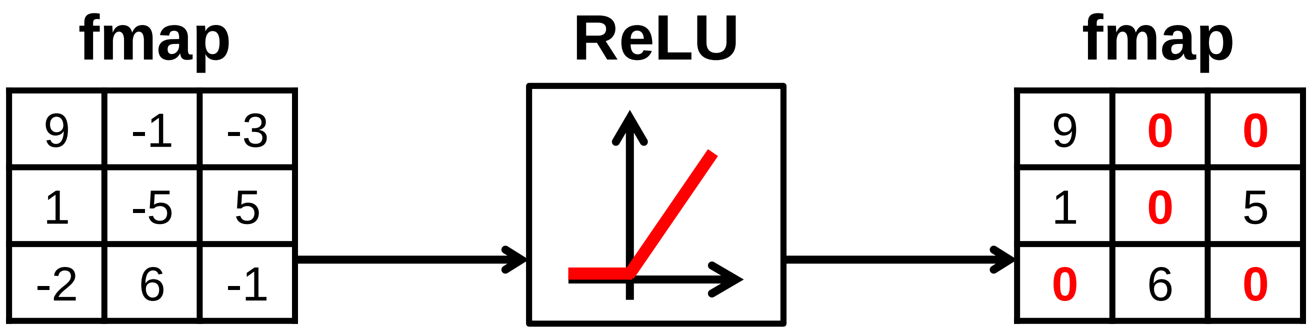


Computation energy scales linearly with the bitwidth of each input

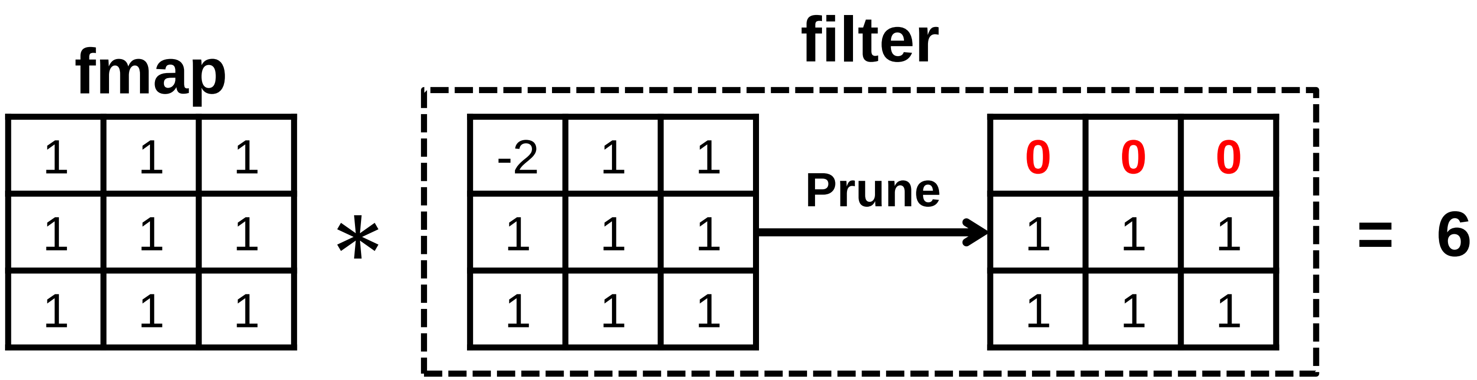


Factor in Sparsity

Apply Non-Linearity **ReLU** on Filtered Image Data

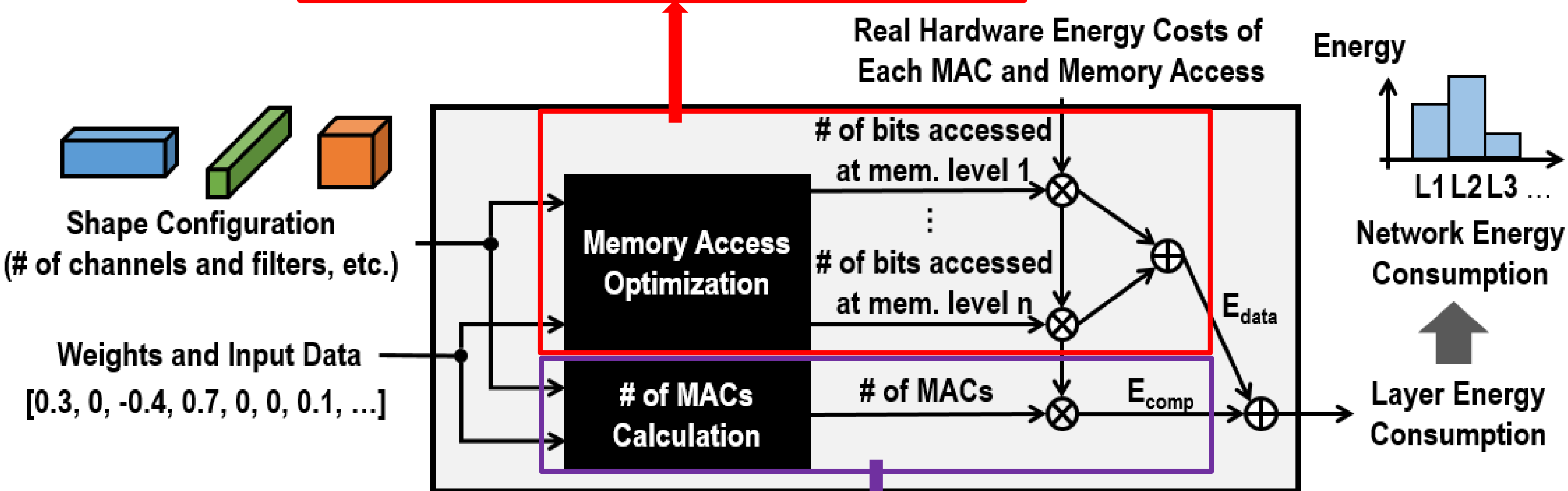


Pruned Network Filters



Factor in Sparsity

Use data compression to reduce the # of bits accessed



Skip the MAC when at least one input is zero

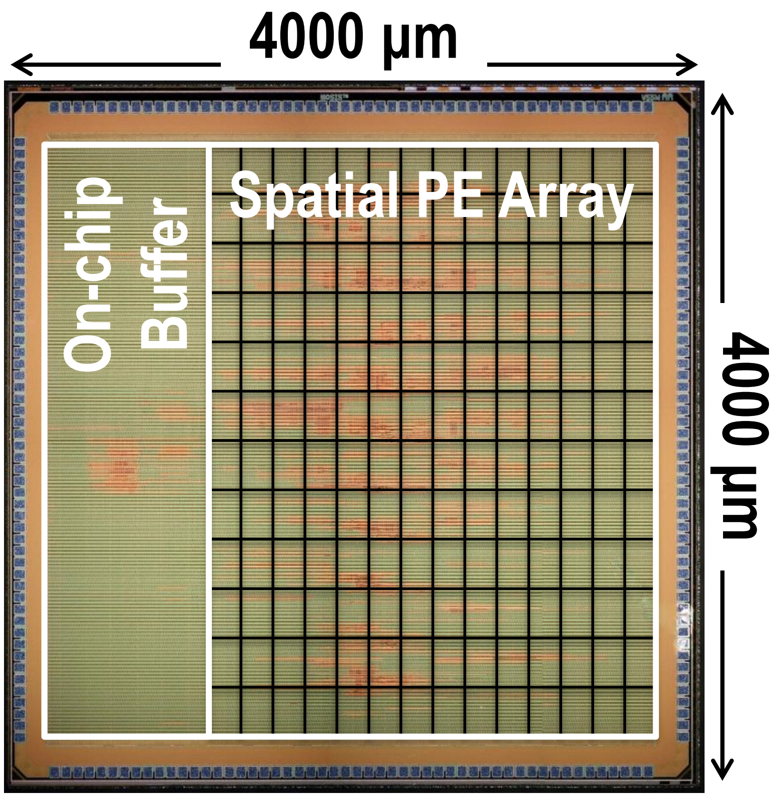
Insights

Example Platform

Eyeriss [*ISSCC*, 2016]

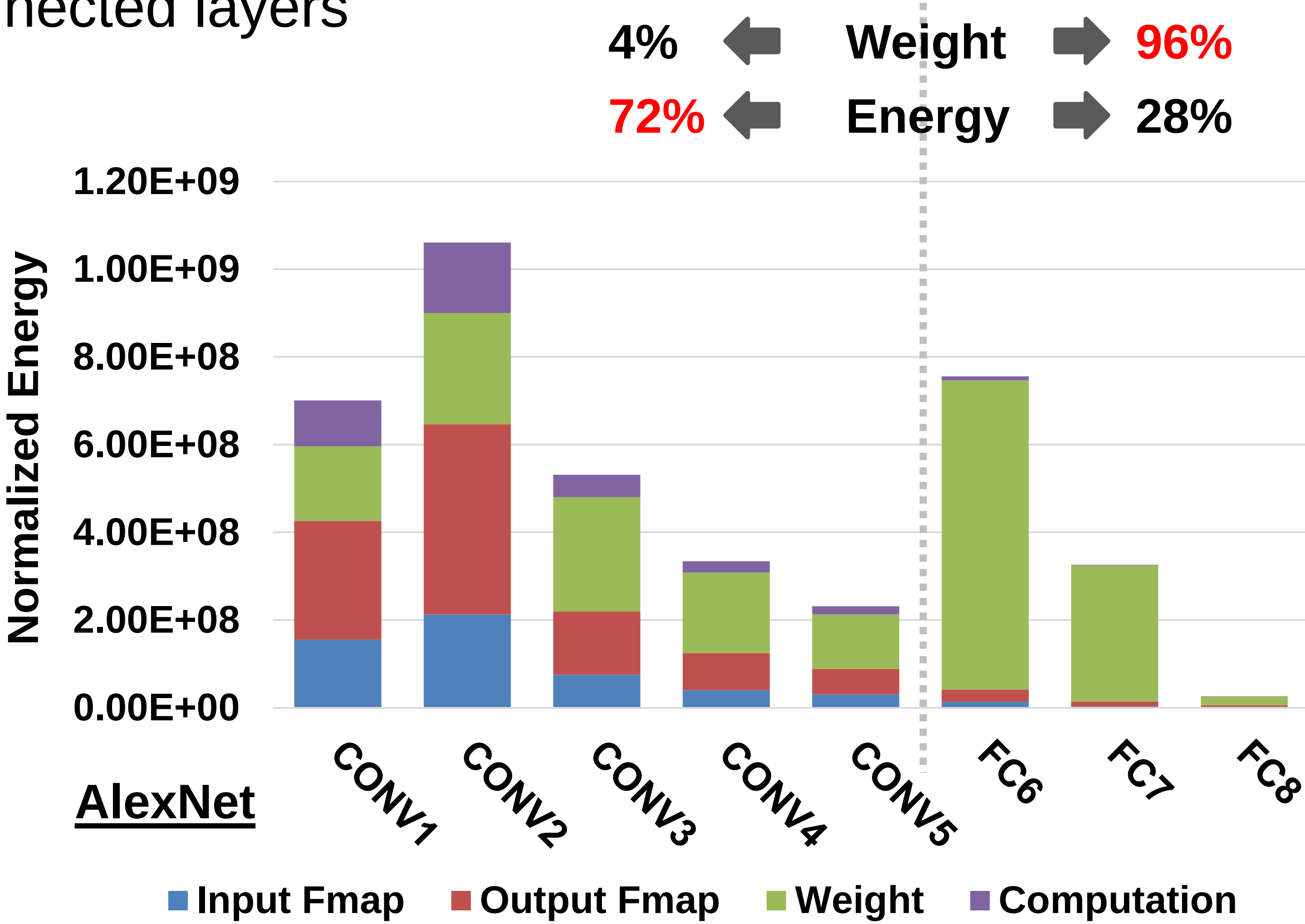
A reconfigurable CNN processor

35 fps @ 278 mW*



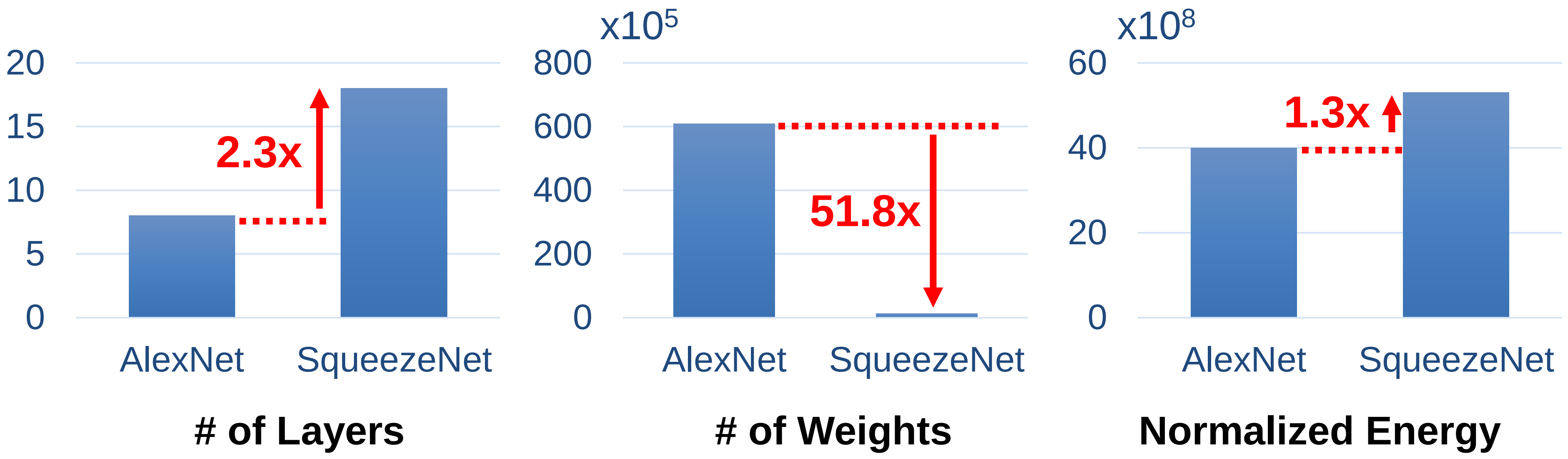
Key Insights

Convolutional layers consume more energy than fully-connected layers



Key Insights

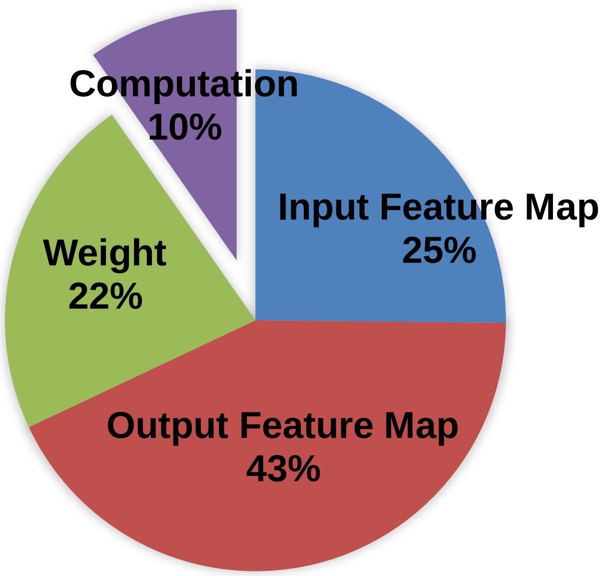
Deeper CNNs with fewer weights do not necessarily consume less energy than shallower CNNs with more weights



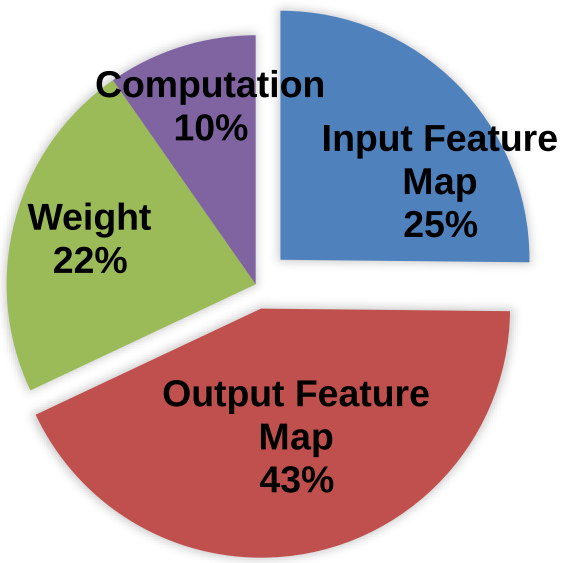
SqueezeNet: F. N. Iandola et al., “SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5MB model size,” arXiv:1602.07360, 2016.

Key Insights

- Data movement is more expensive than computation
- Feature maps need to be taken into account



Computation 10%



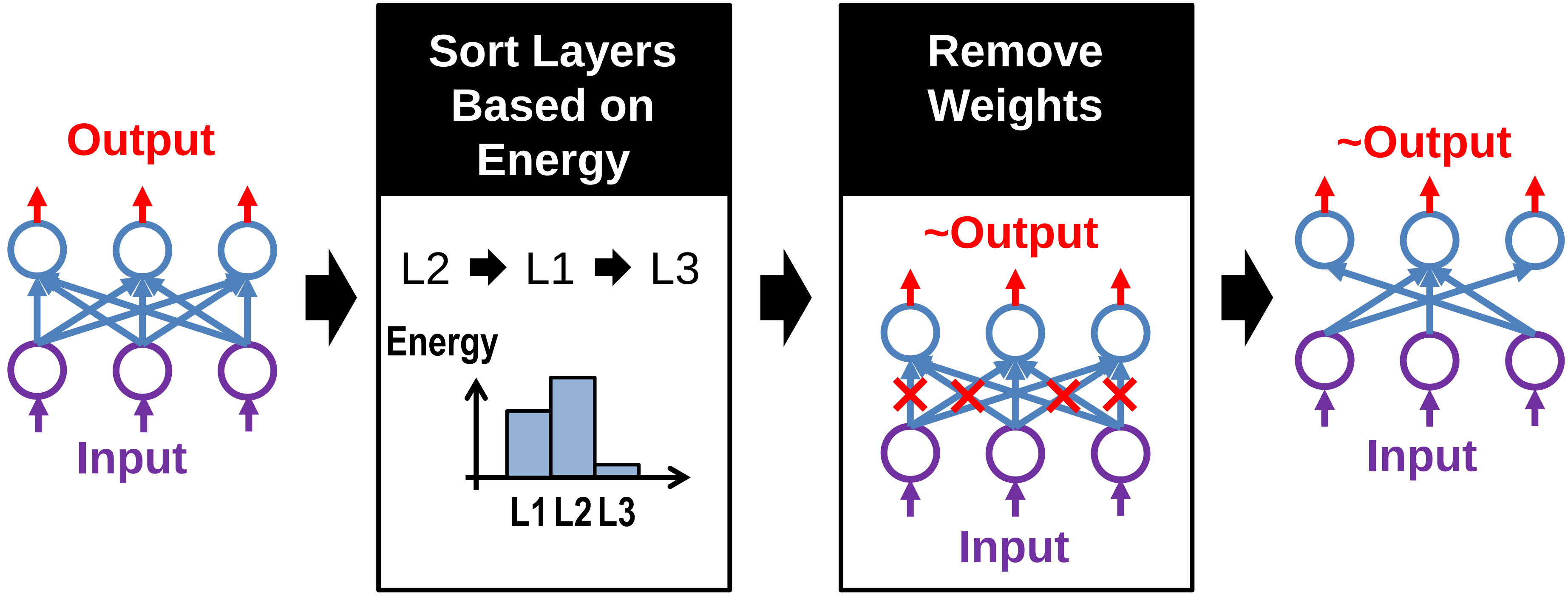
Feature Map 68%

GoogLeNet Energy Breakdown

Application

Energy-Aware Pruning (EAP)

- Use estimated energy to guide the layer-by-layer pruning
- Start from pruning the layers that consume the most of energy



T.-J. Yang et al., “Designing Energy-Efficient Convolutional Neural Networks using Energy-Aware Pruning,” *CVPR*, July 2017.

Energy-Aware Pruning (EAP)

We remove the weights having the **smallest joint impact** on the output instead of the **small magnitude** weights

Magnitude-based Method

$$\underbrace{\begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}}_{\text{fmap}} * \overbrace{\begin{bmatrix} -2 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix} \xrightarrow{\text{Prune}} \begin{bmatrix} -2 & \color{red}{0} & \color{red}{0} \\ \color{red}{0} & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}}^{\text{filter}} = 6 \color{red}{\rightarrow 3}$$

Error: 3

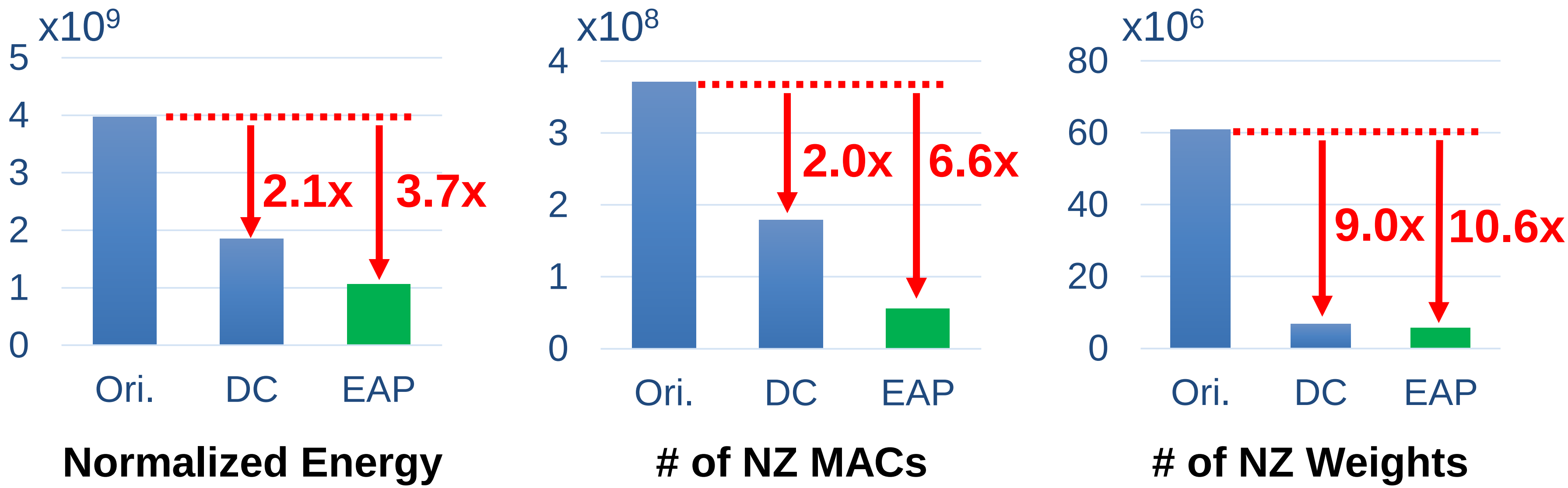
Our Method

$$\underbrace{\begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}}_{\text{fmap}} * \overbrace{\begin{bmatrix} -2 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix} \xrightarrow{\text{Prune}} \begin{bmatrix} \color{red}{0} & \color{red}{0} & \color{red}{0} \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}}^{\text{filter}} = 6$$

Error: 0

Pruned Result Analysis

- EAP reduces AlexNet energy by **3.7x** and outperforms the previous work by **1.7x**
- Energy is more difficult to reduce than # of weights and MACs

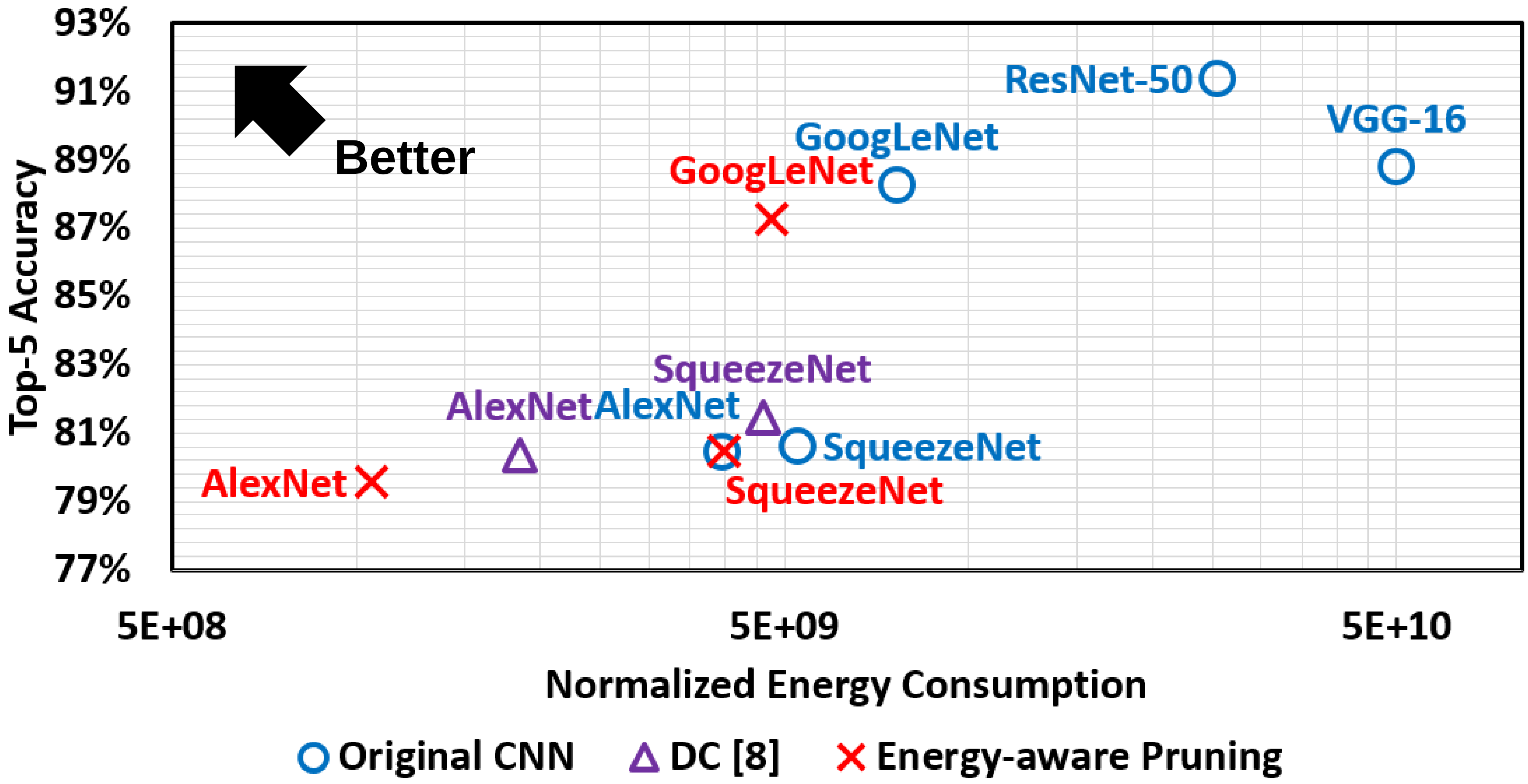


AlexNet

DC: S. Han et al., "Deep Compression: Compressing Deep Neural Networks with Pruning, Trained Quantization and Huffman Coding," in ICLR, 2016.

Network Comparison

- Energy-aware pruning achieves better trade-off



Summary

- We proposed an energy estimation methodology of DNNs based on the architecture, bitwidth and sparsity

- We showed that
  – # of weights and MACs are not good metrics for energy
  – data movement is more expensive than computation
  – feature maps need to be taken into account

- Better accuracy-energy trade-off can be achieved by combining the energy estimation methodology with pruning

Thank You

Learn more about **energy-aware pruning** at
http://eyeriss.mit.edu/energy.html

Learn more about **efficient neural networks** at
https://arxiv.org/abs/1703.09039