Energy-Constrained Compression for Deep Neural Networks via Weighted Sparse Projection and Layer Input Masking

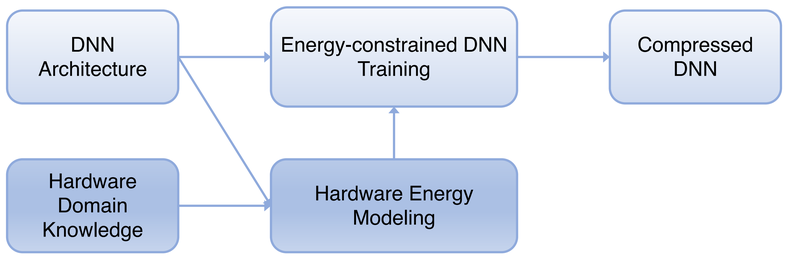
Haichuan Yang[1], Yuhao Zhu[1], Ji Liu[2,1]
[1]University of Rochester, [2]Kwai AI Lab at Seattle

Motivation

- Deep Neural Networks (DNNs) are increasingly deployed in highly energy-constrained environments.
- Compression methods like pruning or quantization are proposed to reduce the redundant parameters.
- Existing work: barely consider the hardware and energy constraint in pruning.
- This work: use the hardware energy model to guide the pruning.

Overview

Overview of the framework:



The overall objective:

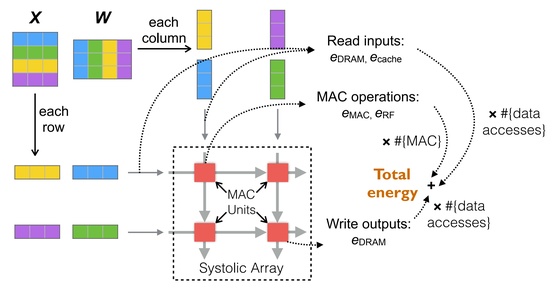
$$\min_{W} \mathcal{L}(W) \quad \text{s.t. } E(W) \le E_{\text{budget}}$$

DNN Parameters; DNN Training Loss; Energy Modeling

Modeling DNN Inference Energy Consumption

Targeted Hardware: systolic-array-based.
Count the hardware operations:

- Data reuse in parallel MACs;
- Zero operands are skipped.



Energy Model: piece-wise linear function $E(W) =$

$$\sum_{i=1}^{L} \alpha^{(i)} \min(k, \|W^{(i)}\|_0) + \beta^{(i)} \max\left(0, \|W^{(i)}\|_0 - k\right) + \gamma^{(i)} \|W^{(i)}\|_0 + \delta$$

cache size; unit energy cost (out-cache); constant overhead; unit energy cost (in-cache); unit energy cost (cache-independent)

Energy-Constrained DNN Training

Basic Idea: Projected-SGD

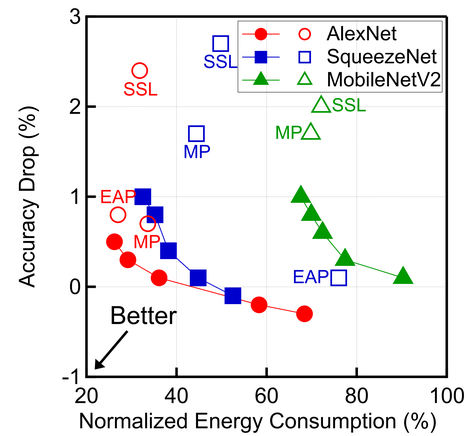
$$W \leftarrow \mathrm{P}\left(W - \eta \nabla \mathcal{L}(W)\right) \quad \text{(SGD)}$$

$$\mathrm{P}(Z) := \operatorname*{argmin}_{W:\, E(W) \le E_{\text{budget}}} \|W - Z\|$$

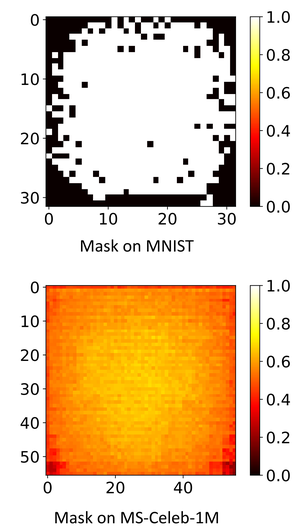
Projection; Theorem 1 ⟺ 0/1 knapsack

Additional trick: design a trainable binary mask to increase the sparsity of the input mask.

Experiment Results



MP: Magnitude-based Pruning;
SSL: Structured Sparsity Learning;
EAP: Energy-Aware Pruning.



DNNs	AlexNet					SqueezeNet				MobileNetV2		
Energy Budget	26%					38%				68%		
Methods	MP	SSL	EAP	NetAdapt	**Ours**	MP	SSL	EAP	**Ours**	MP	SSL	**Ours**
Accuracy Drop	0.7%	2.4%	0.8%	4.4%	**0.5%**	1.7%	2.7%	**0.1%**	0.4%	1.7%	2.0%	**1.0%**
Energy	34%	32%	27%	**26%**	**26%**	44%	50%	76%	**38%**	70%	72%	**68%**
Nonzero Weights Ratio	8%	35%	9%	10%	31%	34%	61%	28%	48%	52%	63%	63%

DNNs@Dataset	LeNet-5@MNIST						MobileNetV2@MS-Celeb-1M		
Energy Budget	17%						60%		
Methods	MP	SSL	NetAdapt	SBP	BC	**Ours**	MP	SSL	**Ours**
Accuracy Drop	1.5%	1.5%	0.6%	1.5%	2.2%	**0.5%**	1.1%	0.7%	**0.2%**
Energy	18%	20%	18%	22%	26%	**17%**	66%	72%	**60%**

SBP: Structured Bayesian Pruning; BC: Bayesian Compression

Conclusion

- Propose an end-to-end energy-aware model compression framework;
- Outperform state-of-the-arts;
- First work to involve the hardware model in model compression.