Towards Energy-efficient and Reliable Machine Learning Accelerators

MHI Scholar Finalist Talk Competition
University of Southern California, Los Angeles

By: Souvik Kundu

Advisors: Peter A. Beerel, Massoud Pedram

September 21, 2020
We are in the Machine Learning (ML) Era
We are in the Machine Learning (ML) Era
We are in the Machine Learning (ML) Era

Cloud computation

- CPUs
- GPUs
- ASICS

Retrained weights

Send sensitive data

Edge computation

Inference

© Souvik Kundu
We are in the Machine Learning (ML) Era

Cloud computation

Edge computation

Retrained weights

Data security issue

Send sensitive data

Train

Inference
We are in the Machine Learning (ML) Era

Cloud computation

Retrained weights

Data security issue

Send sensitive data

Edge computation

Train

Inference

Carbon footprint
We are in the Machine Learning (ML) Era

Cloud computation
- Carbon footprint
- Data security issue
- Reduce Training Complexity

Edge computation
- Train
- Inference
- Reduce Inference Complexity

Reduce Training Complexity
Three Major Thrusts of Our Research

- Algorithmic development
- Hardware capabilities
Three Major Thrusts of Our Research

1. A sparse convolutional neural network (CNN) model

   Algorithmic development

   Hardware capabilities

   Reduction in training energy
Three Major Thrusts of Our Research

1. A sparse convolutional neural network (CNN) model
2. A novel training strategy to ensure the robustness for compressed models

- Algorithmic development
- Hardware capabilities
- Reduction in training energy
- Increased robustness with reduced inference energy

© Souvik Kundu
Three Major Thrusts of Our Research

1. A sparse convolutional neural network (CNN) model
   - Reduction in training energy

2. A novel training strategy to ensure the robustness for compressed models
   - Increased robustness with reduced inference energy

3. A novel compression strategy for event driven deep spiking neural networks (SNNs)
   - Extremely reduced inference energy through event driven computation

© Souvik Kundu
Reducing Training Complexity of CNNs

A CNN for image classification

Basic convolution (CONV) operation

Issues with existing low-complexity models

- Need various types of convolution operation support
- Indexing overhead of channel shuffling

Responsible for majority of the energy consumed
A Pre-defined Sparse CNN

Reduction in trainable parameters
A Pre-defined Sparse CNN

Reduction in trainable parameters

Increasing access energy cost

Off-chip DRAM

On-Chip SRAM

Scratchpad

Compute Unit

Indices

Weights

© Souvik Kundu
A Pre-defined Sparse CNN

Reduction in trainable parameters

Increasing access energy cost

Off-chip DRAM
On-Chip SRAM

Indices
Weights

Repeat the indices
Performance Drop

© Souvik Kundu
A Pre-defined Sparse CNN

Repeat the indices

Dense filters

Increasing access energy cost

Reduction in trainable parameters

Off-chip DRAM

On-Chip SRAM

Scratchpad

Compute Unit

Indices

Weights

Improve performance

Reuse indices

Off-chip weights

On-chip weights
A Pre-defined Sparse CNN

- **Dense filters**
- **Off-chip DRAM**
- **On-Chip SRAM**
- **Scratchpad**
- **Compute Unit**

**Reduction in trainable parameters**

**Increasing access energy cost**

**Reuse indices**

**Compute complexity (FLOPs)**:
- ShuffleNet: 0.39G
- MobileNetV2: 0.38G
- Ours: 0.27G

**Classification Accuracy %**:
- ShuffleNet: 55.7%
- MobileNetV2: 56%
- Ours: 59%

*Results on Tiny-ImageNet (top-1) where the proposed model has similar or lesser parameter compared to the other two. All trained with same hyper parameters.*
Robustness is a Growing Concern:
Robustness is a Growing Concern:

```
STOP  ->  Stop  
        ->  Tree  
        ->  House

Take knowledge from model

Attacker
```
Robustness is a Growing Concern:

Robustness is model performance against these perturbed input

Add perturbation to input

Take knowledge from model

Attacker

A life-threatening consequence

Stop
Tree
House
Bird

© Souvik Kundu
Adversarial Training Likes More Weights

\[ \chi + \varepsilon \times \text{sgn}(\nabla_x J(g(x; \theta), t)) = \hat{x} \]
Adversarial Training Likes More Weights

Number of weights having non-negligible magnitudes increases when we train the model with adversarial as well as clean image.
Number of weights having non-negligible magnitudes increases when we train the model with adversarial as well as clean image.

Robust pruning is challenging
Prior Art Approaches: All Iterative

After standard adversarial training
Prior Art Approaches: All Iterative

After standard adversarial training

Iterative Compression

© Souvik Kundu
Prior Art Approaches: All Iterative

Proper tuning of per-layer pruning for better performance is a tedious job.

We use the hidden information of the network to find layer significance: \( \frac{\partial \text{Loss}}{\partial \text{Weight}} \)

momentum
Our Unified Robust Compression

- Calculate momentum distribution per layer
- Prune fraction of smallest weights from each layer
- Redistribute edges according to weights having larger momentums

*Based on results evaluated with VGG16 and ResNet18 on CIFAR datasets.*
Our Unified Robust Compression

- Calculate momentum distribution per layer
- Prune fraction of smallest weights from each layer
- Redistribute edges according to weights having larger momentums

**Loss**
- Loss\textsubscript{clean}
- Loss\textsubscript{adv}

**Regularizer\textsubscript{conver}**
- Better convergence
- Sparse weight-update

Increased inference compute efficiency up to 50% compared to the currently existing approaches*

*Based on results evaluated with VGG16 and ResNet18 on CIFAR datasets.
Extension to Support Channel Pruning

CONV 1 → CONV 2 → CONV 3 → Flatten

Linear layers
Extension to Support Channel Pruning

CONV 1  CONV 2  CONV 3  Flatten

Linear layers

© Souvik Kundu
Extension to Support Channel Pruning

Effective model shrinks

Potential inference speed-up up to 10x

*Based on results evaluated with VGG16 and ResNet18 on CIFAR datasets.
Thinking Beyond Conventional Computation

Analog input driven compute

Artificial neural network
Thinking Beyond Conventional Computation

Analog input driven compute

Artificial neural network

Spike based event-driven compute

Spiking neural network

Synaptic neuron

Synaptic weight

Yes! a spike!

© Souvik Kundu
Thinking Beyond Conventional Computation

Artificial neural network

Analog input driven compute

Spike based event-driven compute

Spiking neural network

Low-power compute with suitable hardware

Need to look for longer duration to see better

Notion of time

Yes! a spike!

Synaptic weight

Synaptic neuron
Deep SNNs: Beauty and the Beast!

- **Synaptic neuron**
- **Integrate and fire**
- **Check if crossing a threshold**
Deep SNNs: Beauty and the Beast!

Integrate and fire

Check if crossing a threshold

No Multiply-add! Only accumulate
Deep SNNs: Beauty and the Beast!

Integrate and fire

Synaptic neuron

Check if crossing a threshold

No Multiply-add! Only accumulate

 Existing approach

Constrained ANN training

ANN-to-SNN Conversion

SNN inference

Our focus

Training time similar to ANN

Increased spiking activity over time

Accuracy similar to ANN

Increased inference energy

© Souvik Kundu
Compression for Brain-inspired Computing!

- A SNN conversion friendly ANN
- Reduced weight ANN
- Spikethrift: reduced spiking activity SNN

Potential for extreme energy-efficient compressed deep SNNs

No batch-norm
No bias
No max-pool

© Souvik Kundu
Compression for Brain-inspired Computing!

- **A SNN conversion friendly ANN**
- **Non-iterative compression**
- **Reduced weight ANN**
- **Spikethrift: reduced spiking activity SNN**

Potential for extreme energy-efficient compressed deep SNNs

Training suffers from convergence issue

No batch-norm
No bias
No max-pool

Apply ANN-to-SNN conversion

© Souvik Kundu

USC Viterbi
School of Engineering
Compression via Brain-inspired Learning!

Potential for extreme energy-efficient compressed deep SNNs

A SNN conversion friendly ANN

Reduced weight ANN

Spikethrift: reduced spiking activity SNN

Let us use someone to guide this fellow

A pre-trained unpruned meta model

No batch-norm

No bias

No max-pool

non-iterative compression

Apply ANN-to-SNN conversion

Compared to ANN with similar parameters, Inference compute energy can reduce up to 3x*

*Based on initial results on CIFAR datasets with VGG model

© Souvik Kundu
Summary

Reduce training energy through a novel convolution-based model.

Reduce inference energy and retain robustness through a unified training via a comprehensive loss.

A guided compression strategy for event-driven SNN to yield extreme energy-efficient models.
Thanks to all ...
QUESTIONS