Towards Energy-efficient and Reliable Machine Learning Accelerators

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

By:

Souvik Kundu

Advisors:

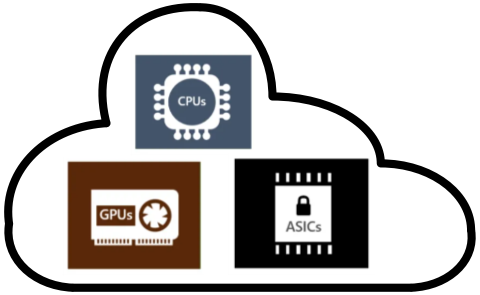
Peter A. Beerel

Massoud Pedram

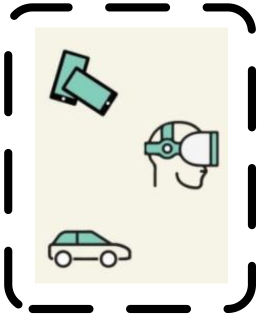
September 21, 2020

We are in the Machine Learning (ML) Era

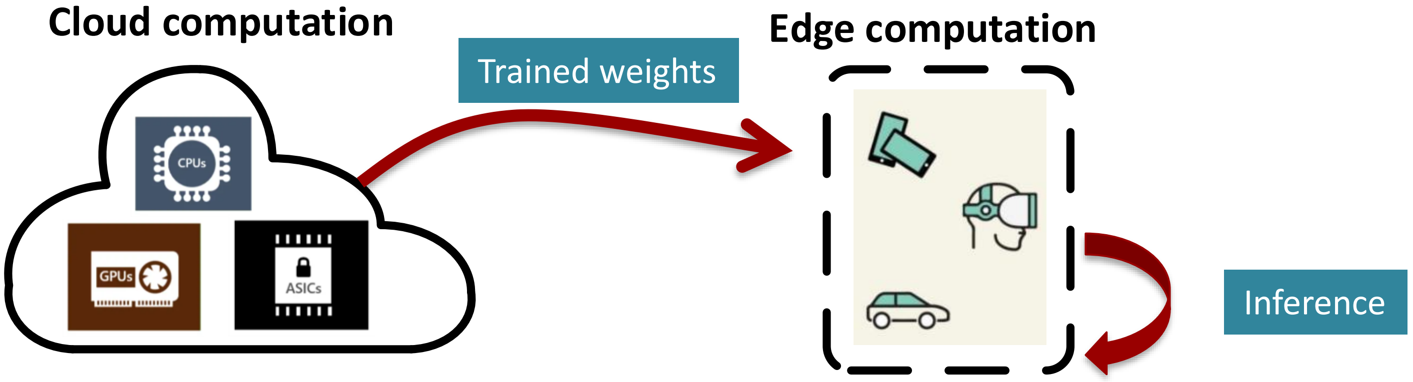
Cloud computation



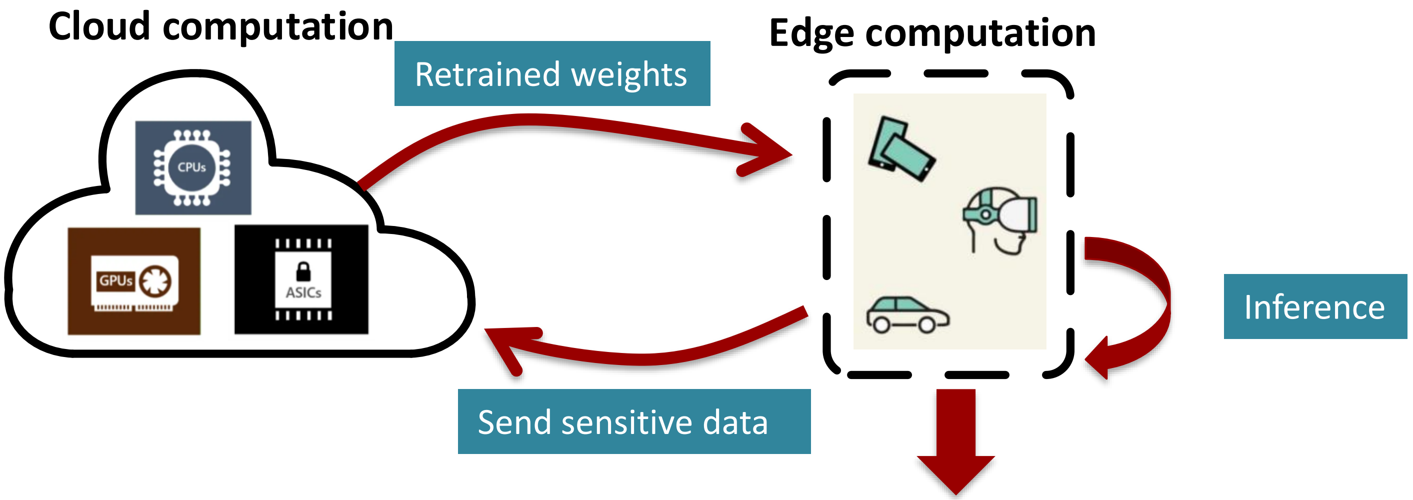
Edge computation



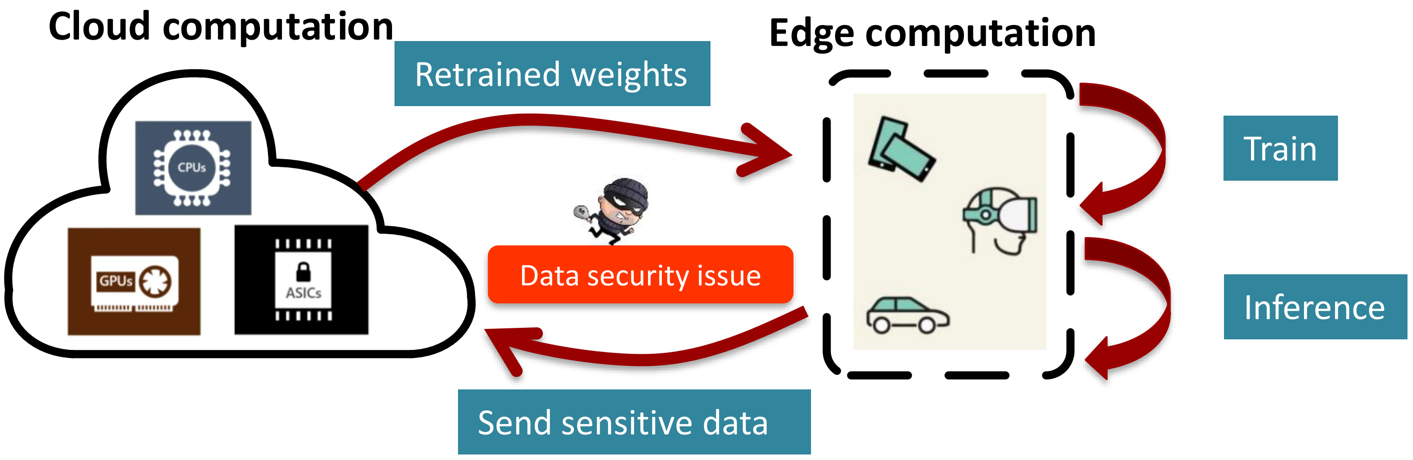
We are in the Machine Learning (ML) Era



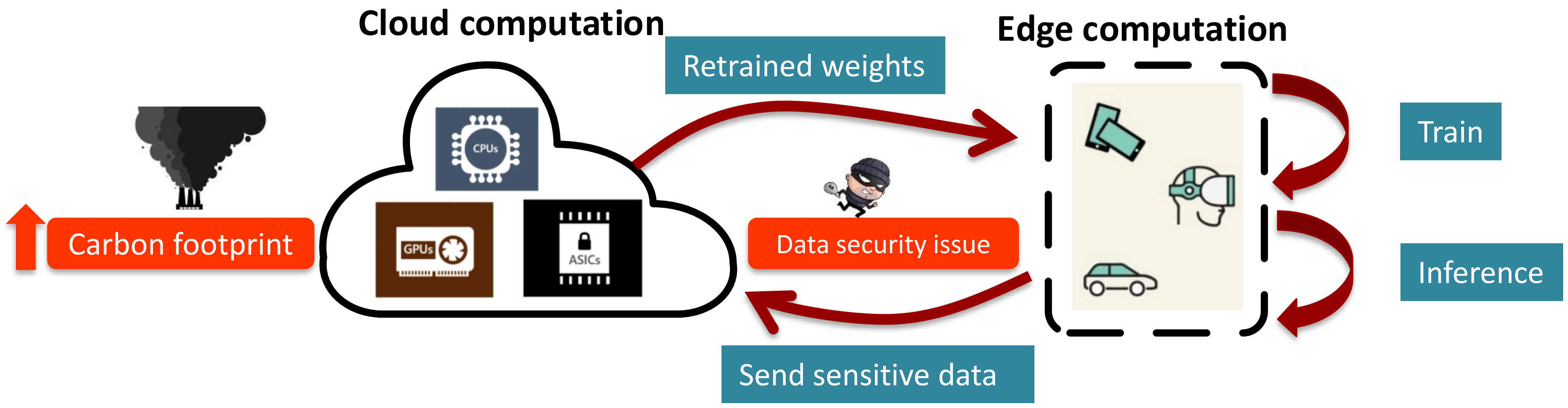
We are in the Machine Learning (ML) Era



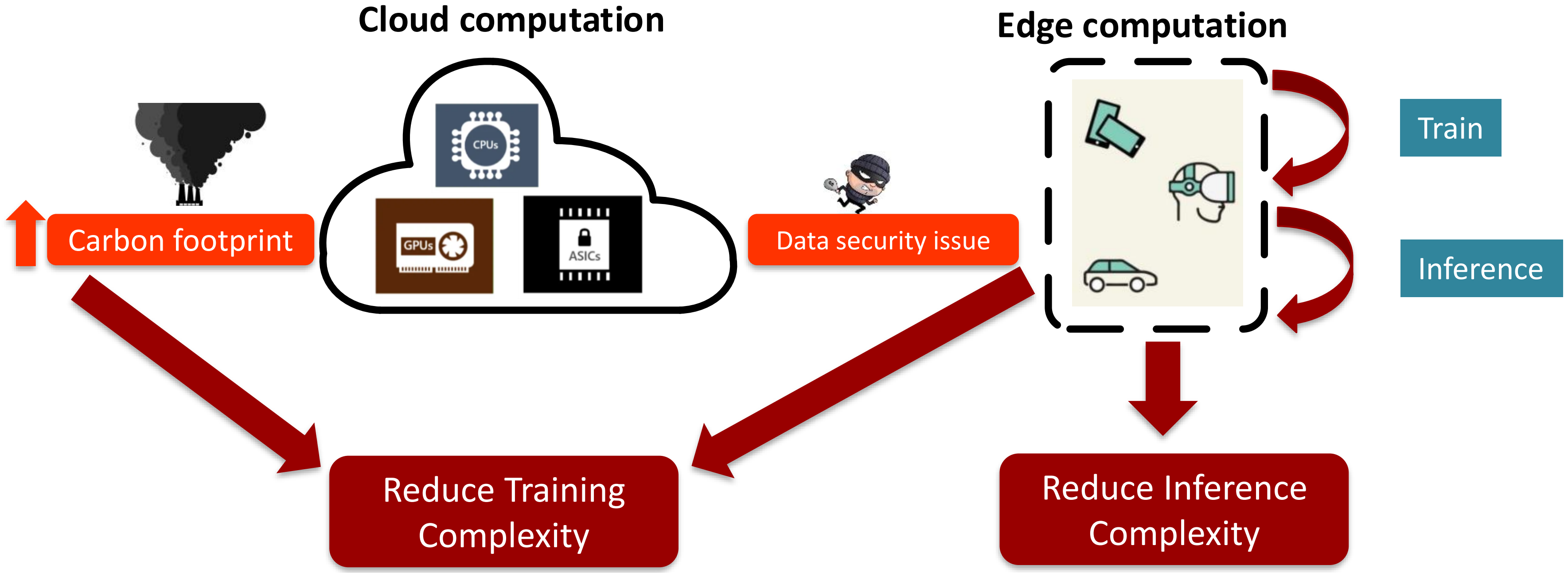
We are in the Machine Learning (ML) Era



We are in the Machine Learning (ML) Era



We are in the Machine Learning (ML) Era



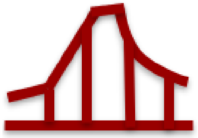
Three Major Thrusts of Our Research

Algorithmic development

Hardware capabilities

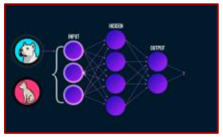
Three Major Thrusts of Our Research

Algorithmic development



Hardware capabilities

1 A sparse convolutional neural network (CNN) model



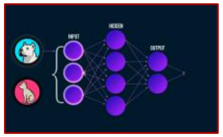
Reduction in training energy

Three Major Thrusts of Our Research

Algorithmic development

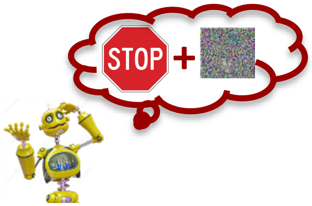
Hardware capabilities

1 A sparse convolutional neural network (CNN) model



Reduction in training energy

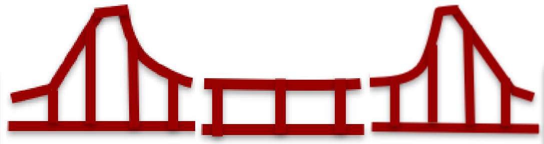
2 A novel training strategy to ensure the robustness for compressed models



Increased robustness with reduced inference energy

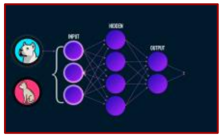
Three Major Thrusts of Our Research

Algorithmic development



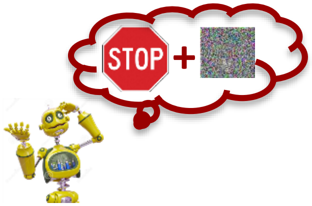
Hardware capabilities

1 A sparse convolutional neural network (CNN) model



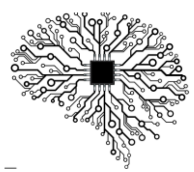
Reduction in training energy

2 A novel training strategy to ensure the robustness for compressed models



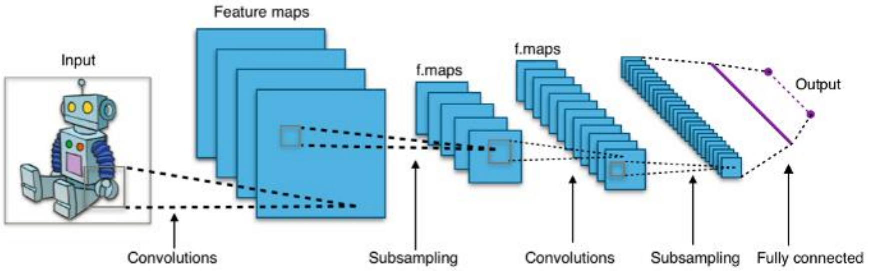
Increased robustness with reduced inference energy

3 A novel compression strategy for event driven deep spiking neural networks (SNNs)

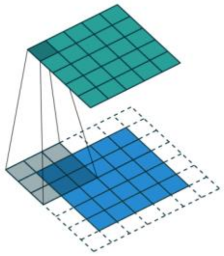


Extremely reduced inference energy through event driven computation

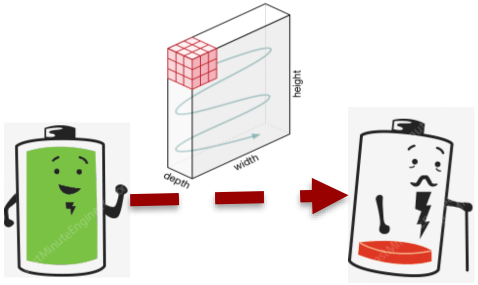
Reducing Training Complexity of CNNs



A CNN for image classification



Basic convolution (CONV) operation

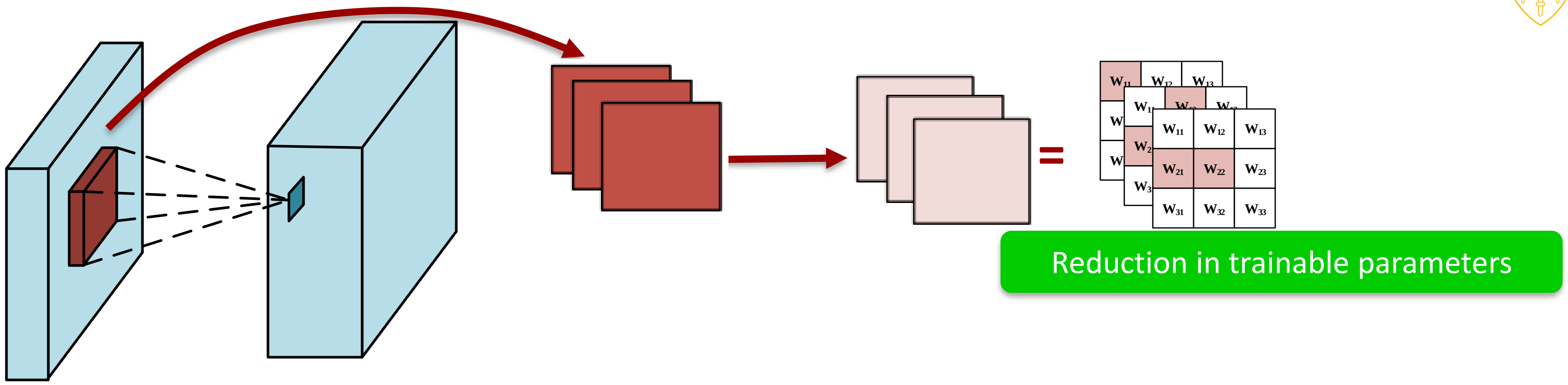


Responsible for majority of the energy consumed

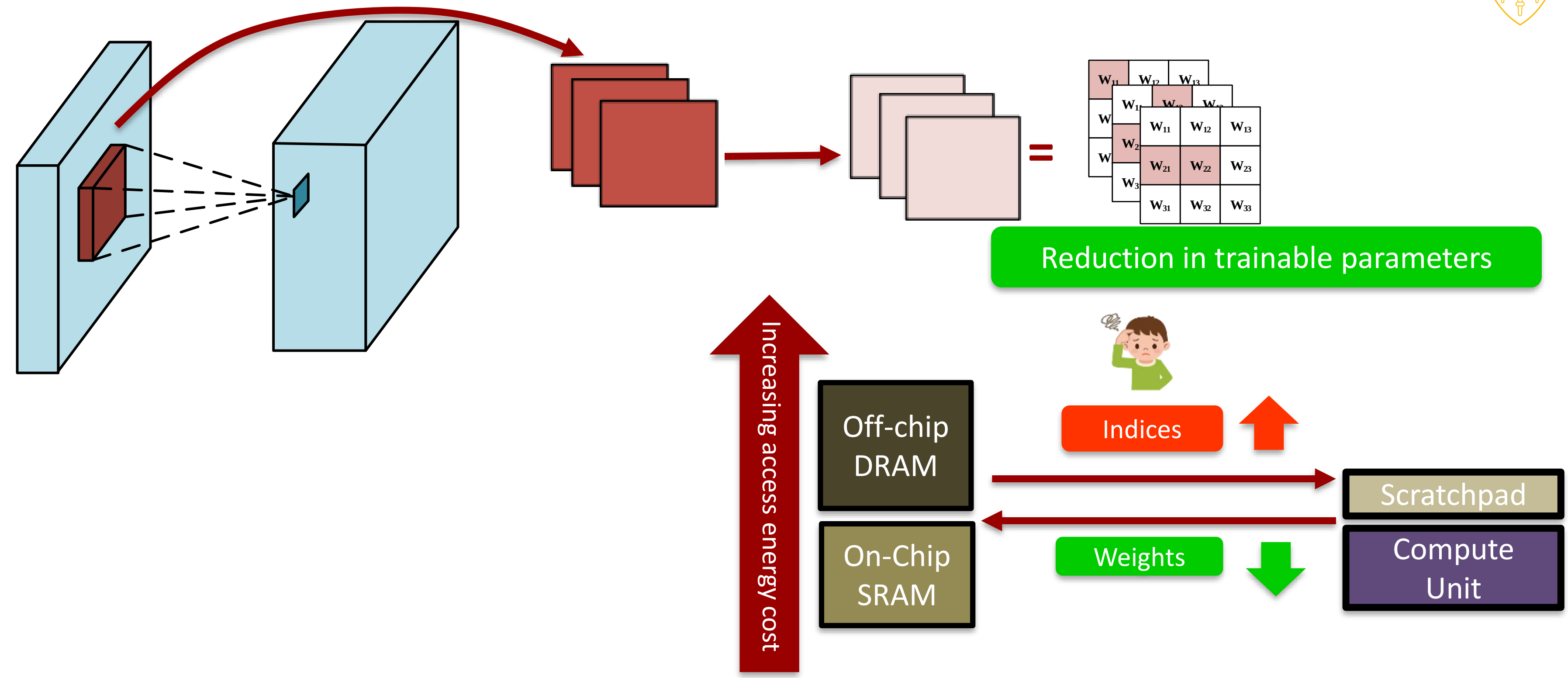
Issues with existing low-complexity models

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

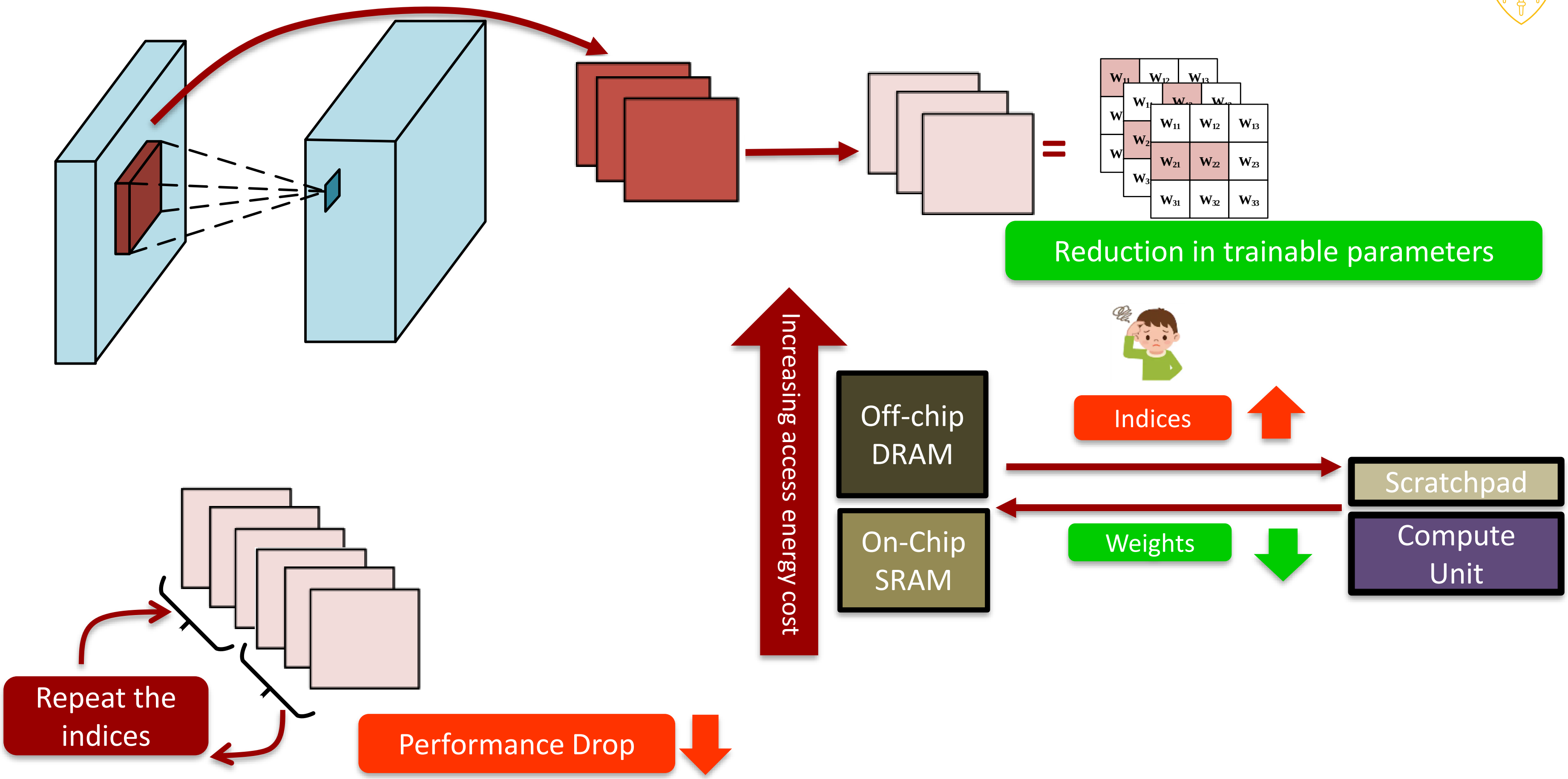
A Pre-defined Sparse CNN



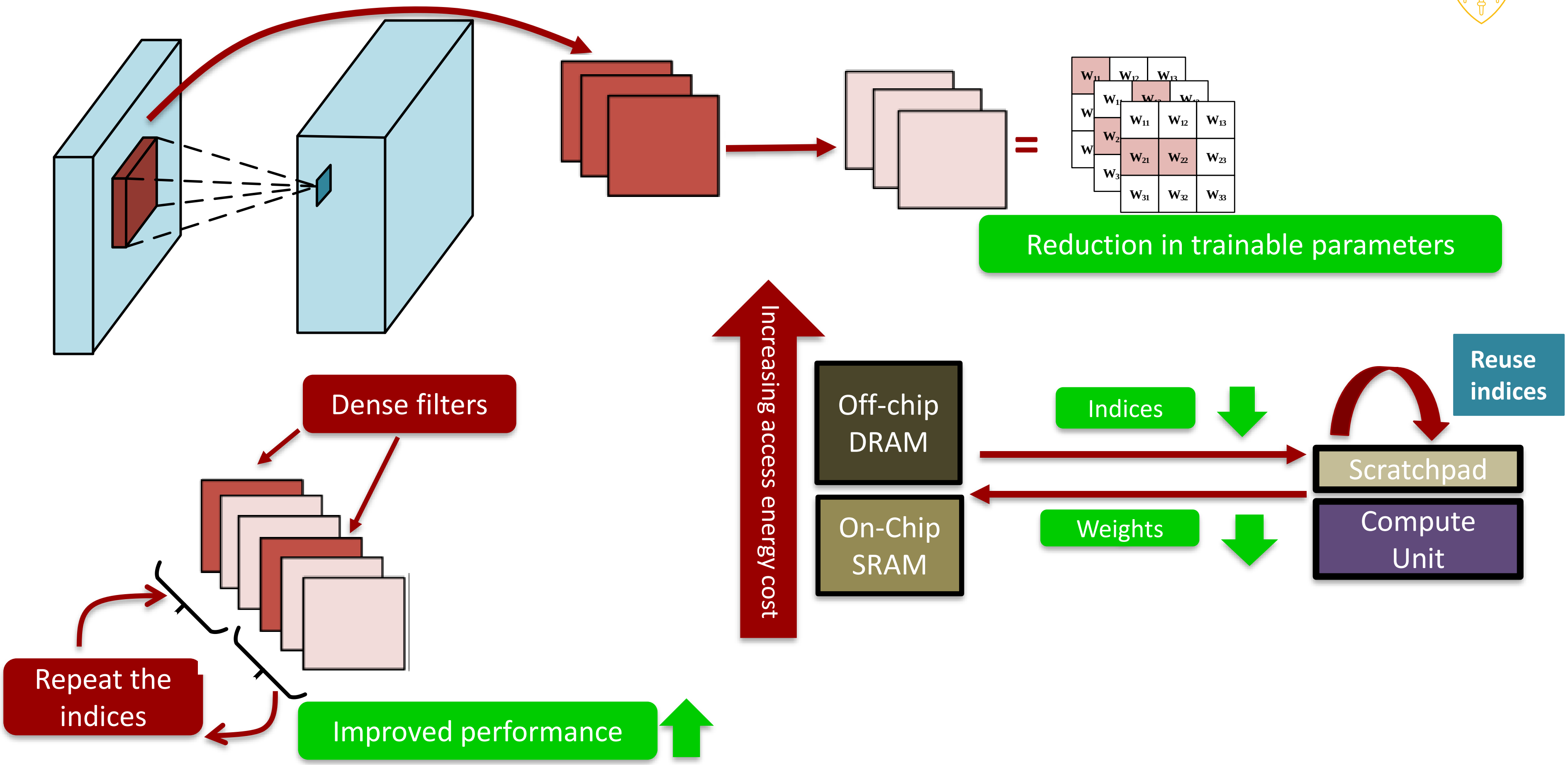
A Pre-defined Sparse CNN



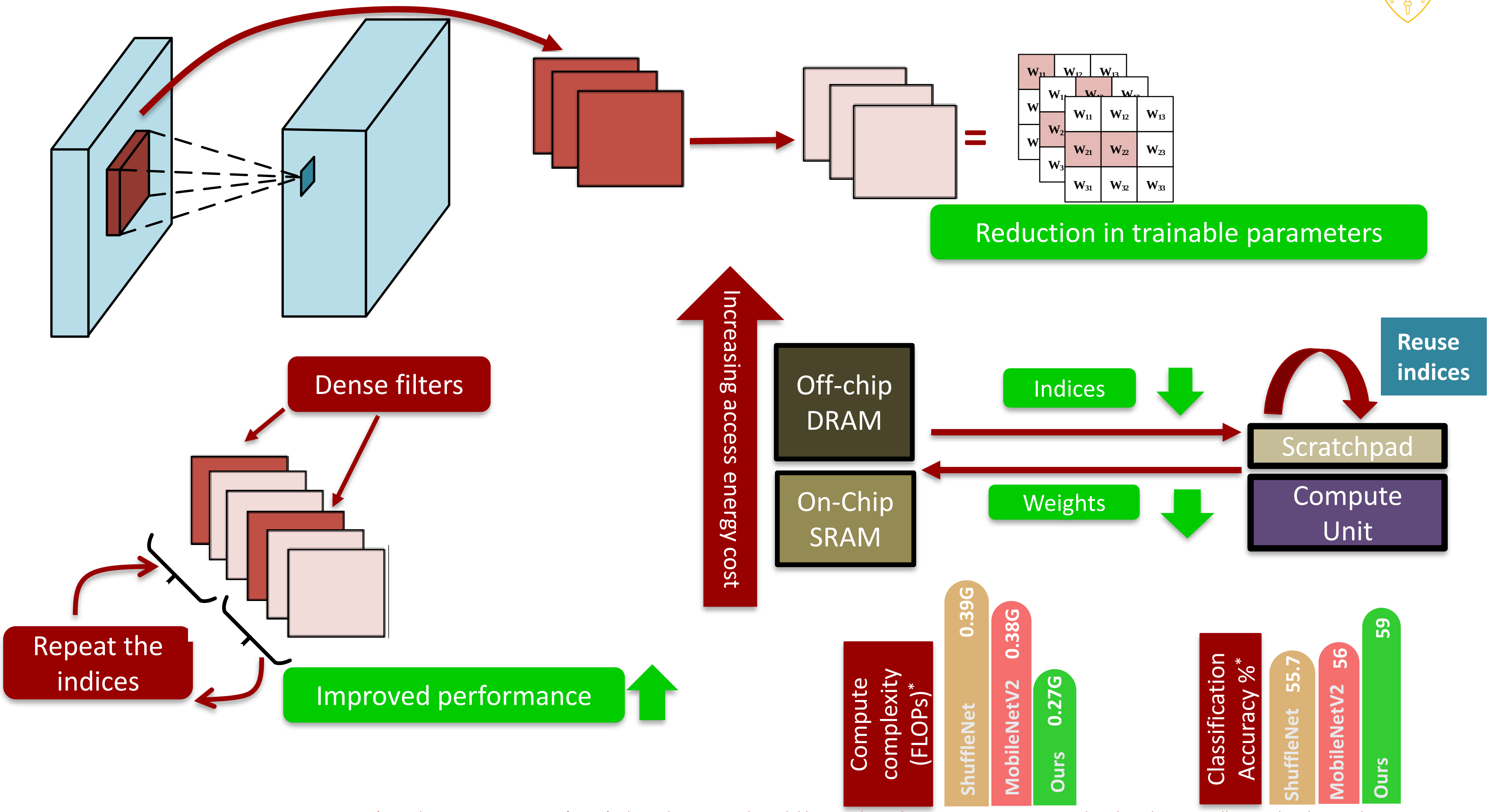
A Pre-defined Sparse CNN



A Pre-defined Sparse CNN

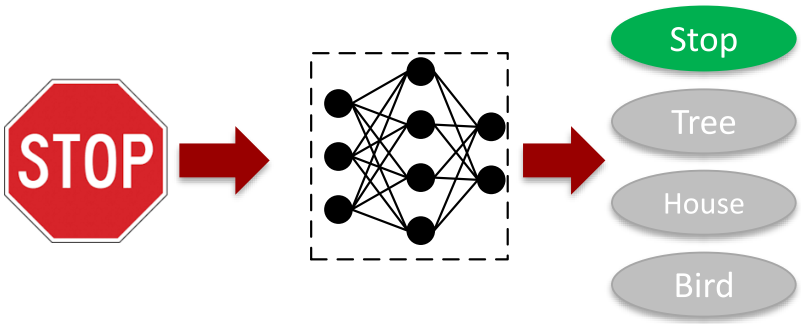


A Pre-defined Sparse CNN

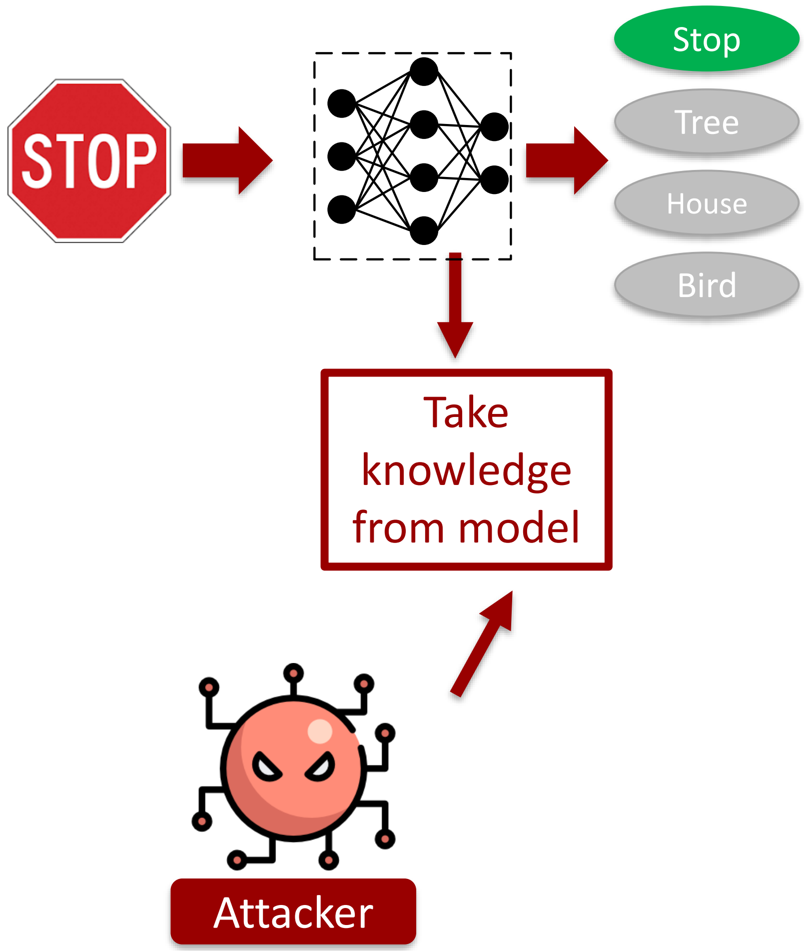


** Results on Tiny-ImageNet (top-1) where the proposed model has similar or lesser parameter compared to the other two. All trained with same hyper parameters.*

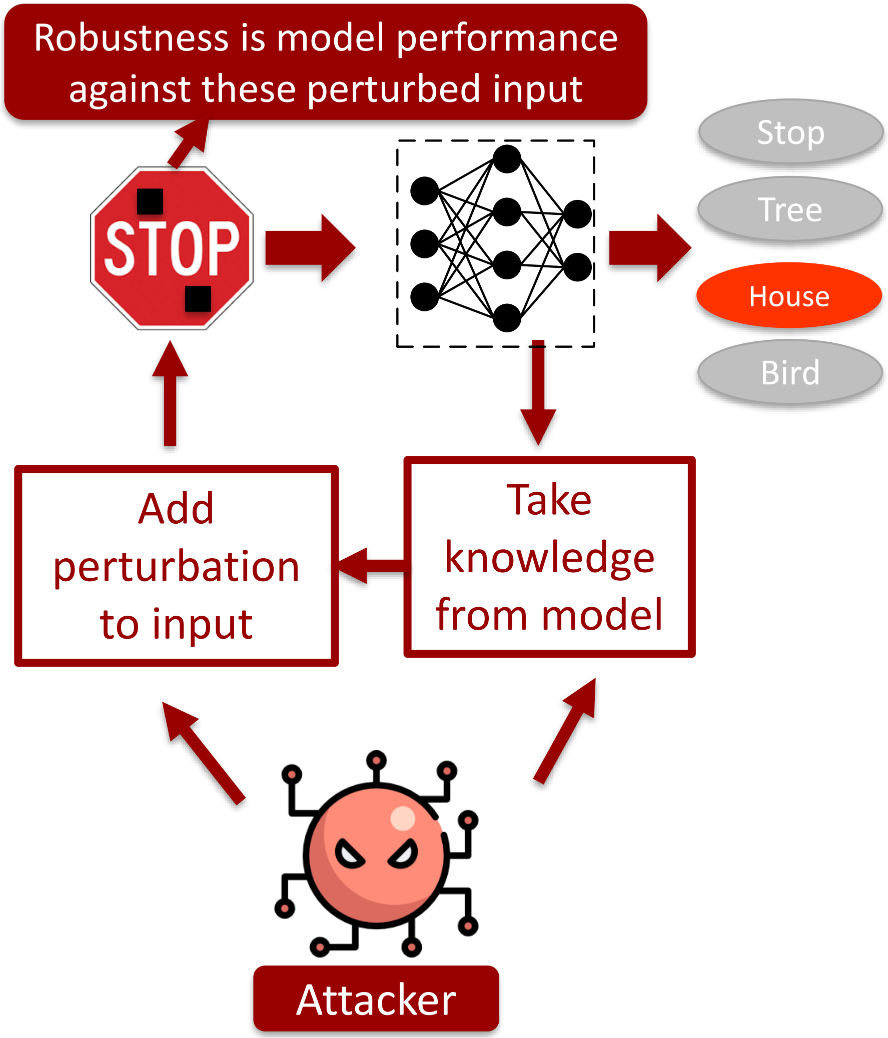
Robustness is a Growing Concern :



Robustness is a Growing Concern :

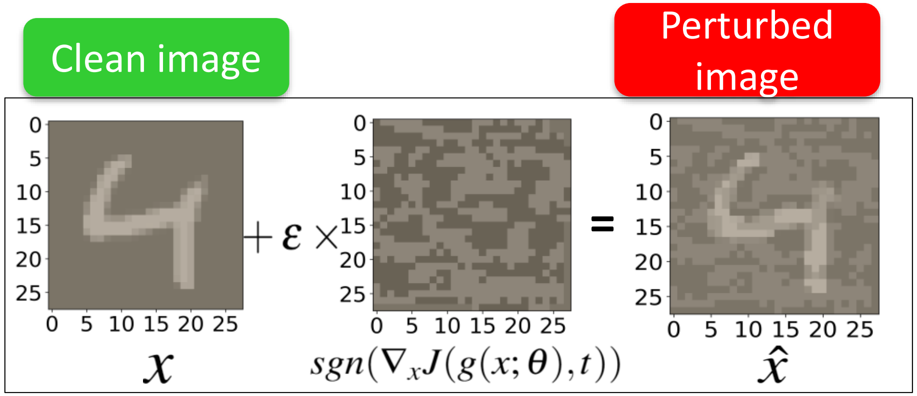


Robustness is a Growing Concern :

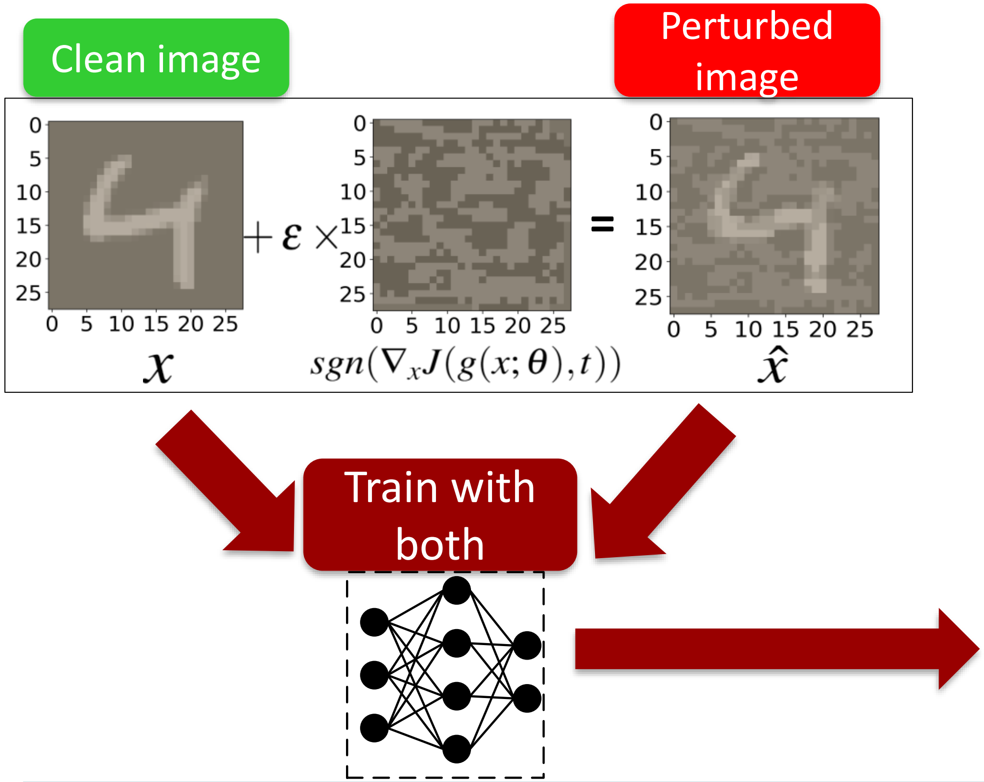


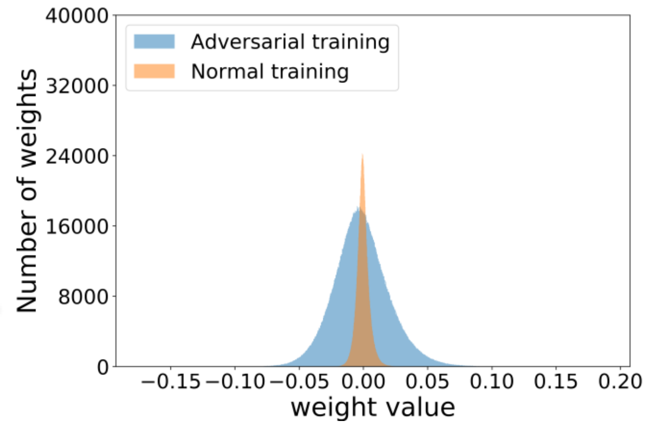
A life-threatening consequence

Adversarial Training Likes More Weights



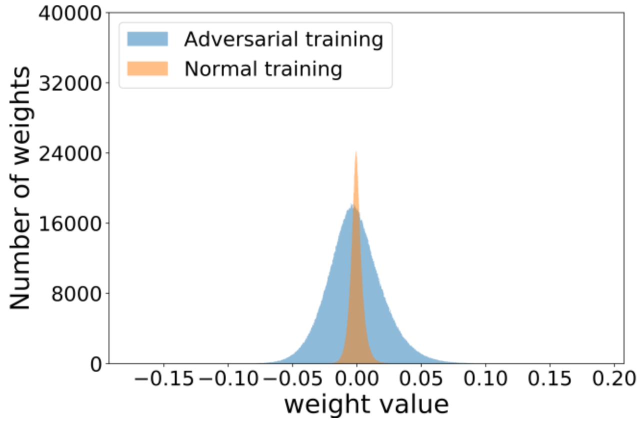
Adversarial Training Likes More Weights





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

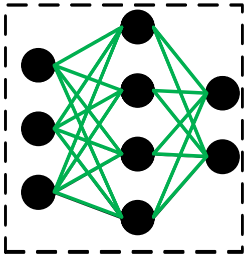
Adversarial Training Likes More Weights



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

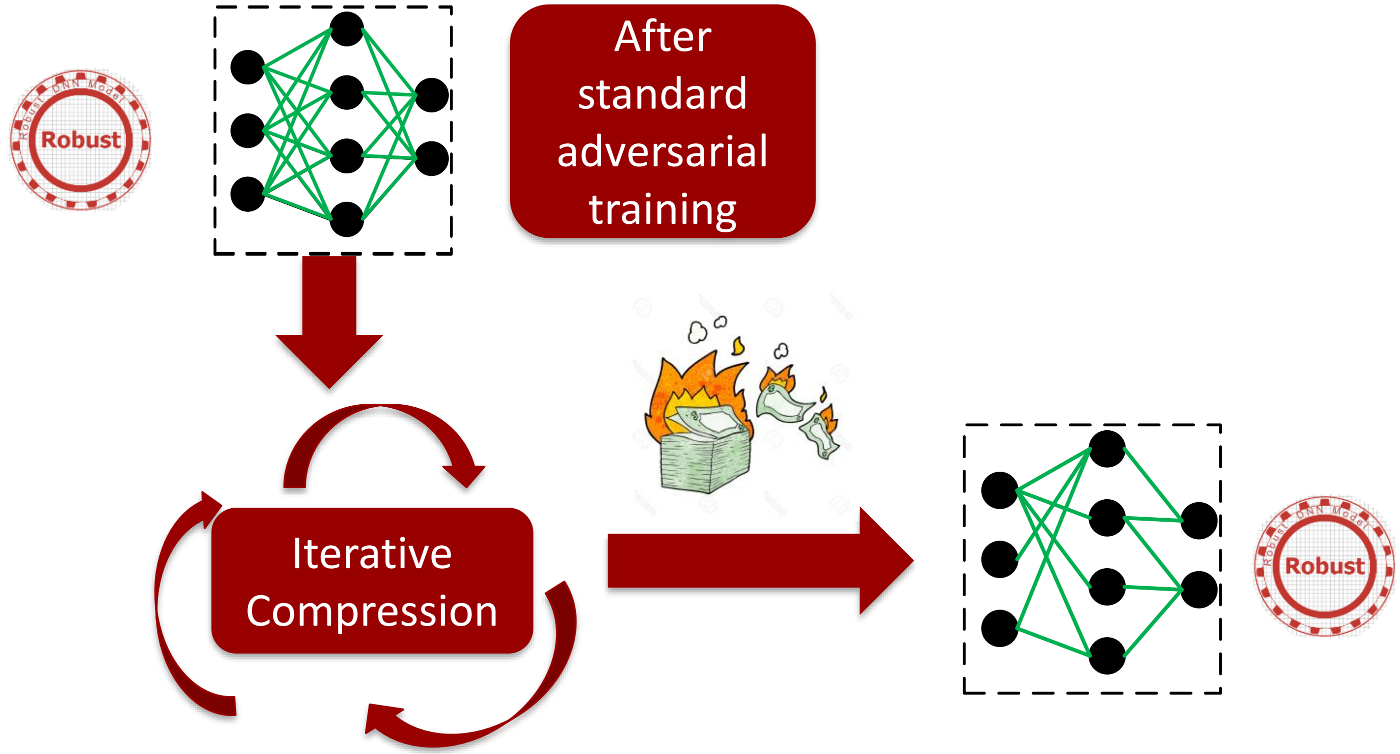
Robust pruning is challenging

Prior Art Approaches: All Iterative



After standard adversarial training

Prior Art Approaches: All Iterative



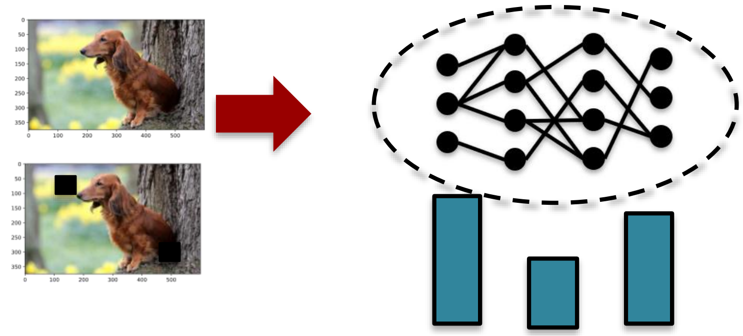
Prior Art Approaches: All Iterative

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

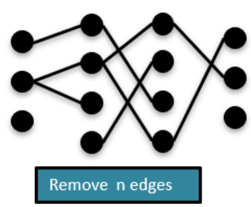
We use the hidden information of the network to find layer significance: $\frac{\partial(Loss)}{\partial(Weight)}$

momentum

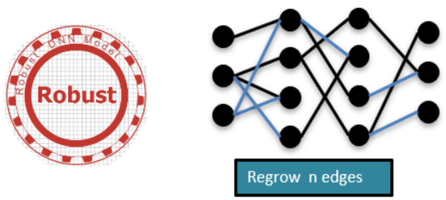
Our Unified Robust Compression



Calculate momentum distribution per layer



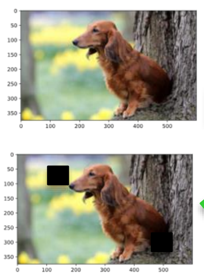
Prune fraction of smallest weights from each layer



Redistribute edges according to weights having larger momentums

*Based on results evaluated with VGG16 and ResNet18 on CIFAR datasets.

Our Unified Robust Compression



Loss

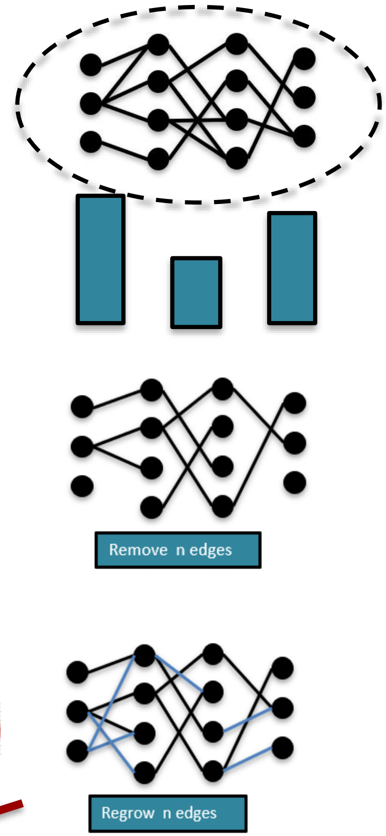
$Loss_{clean}$

$Loss_{adv}$

$Regularizer_{conver}$

Better convergence

Sparse weight-update



Calculate momentum distribution per layer

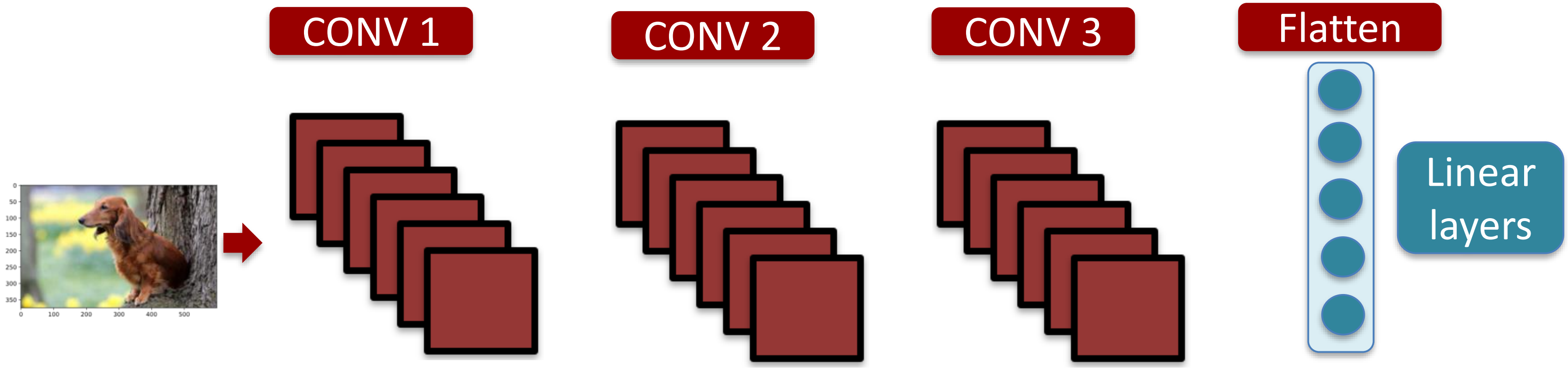
Prune fraction of smallest weights from each layer

Redistribute edges according to weights having larger momentums

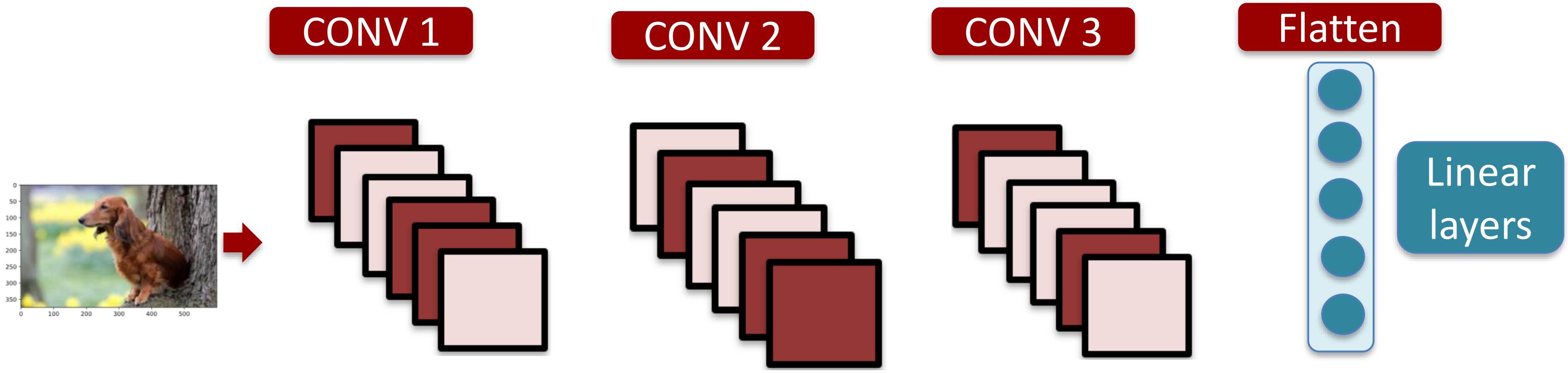
Increased inference compute efficiency up to 50% compared to the currently existing approaches*

*Based on results evaluated with VGG16 and ResNet18 on CIFAR datasets.

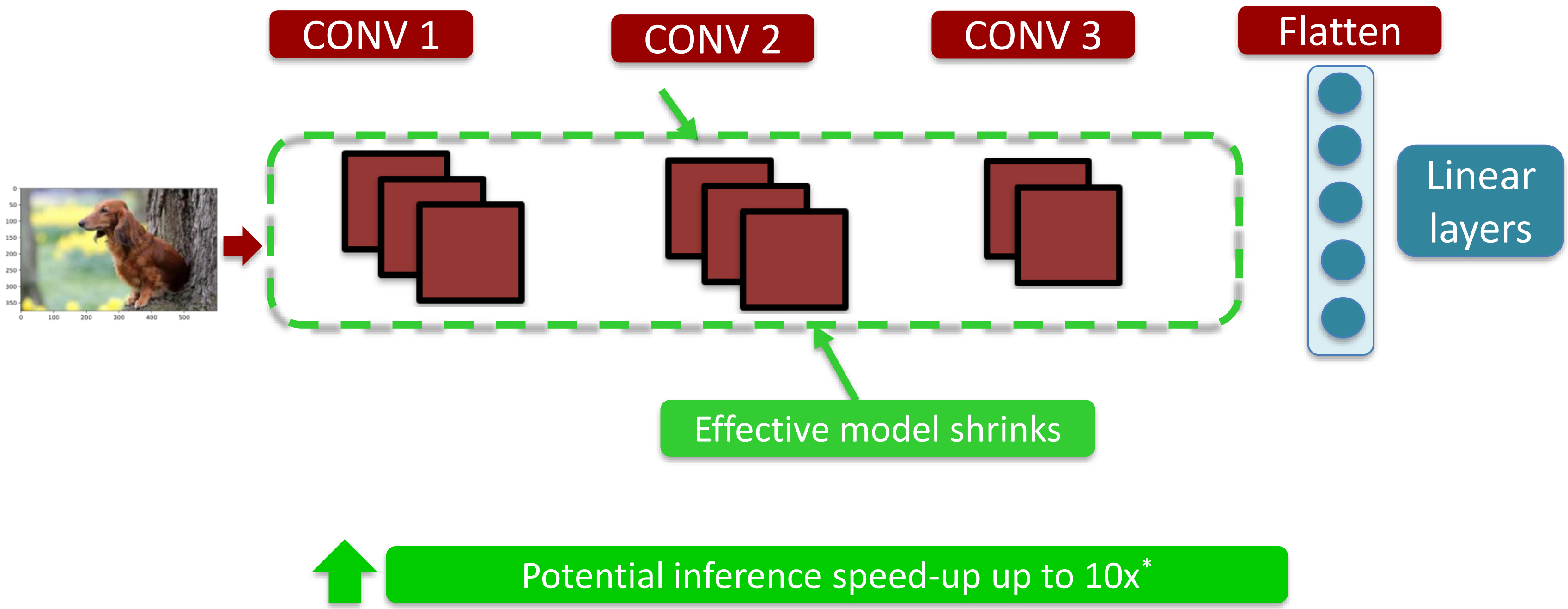
Extension to Support Channel Pruning



Extension to Support Channel Pruning

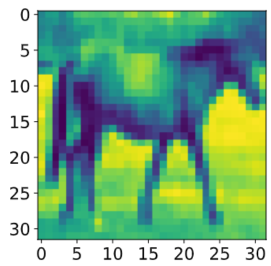


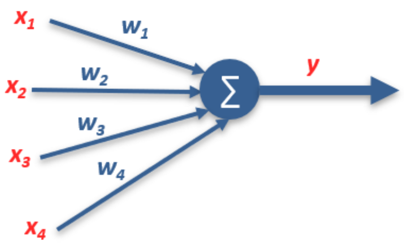
Extension to Support Channel Pruning



*Based on results evaluated with VGG16 and ResNet18 on CIFAR datasets.

Thinking Beyond Conventional Computation

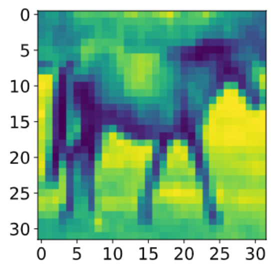


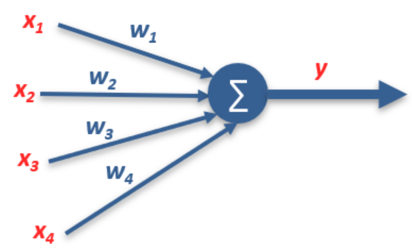


Artificial neural network

Analog input driven compute

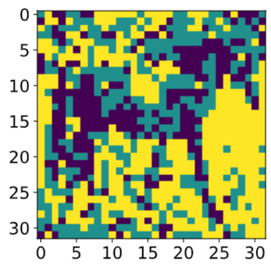
Thinking Beyond Conventional Computation

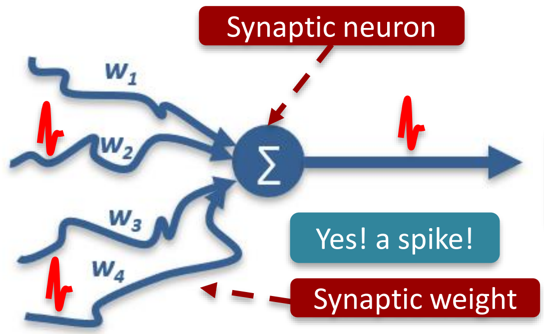




Artificial neural network

Analog input driven compute

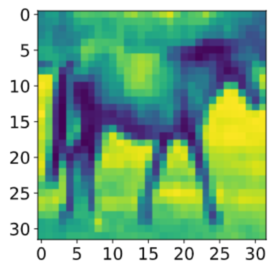


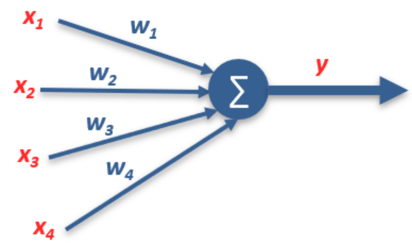


Spiking neural network

Spike based event-driven compute

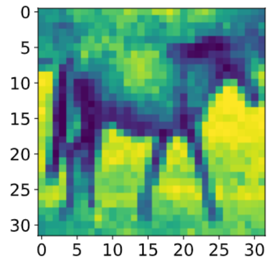
Thinking Beyond Conventional Computation

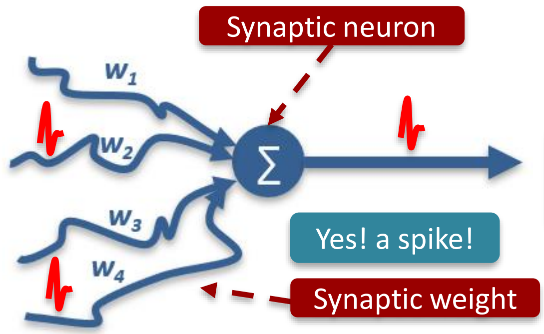




Artificial neural network

Analog input driven compute





Spiking neural network

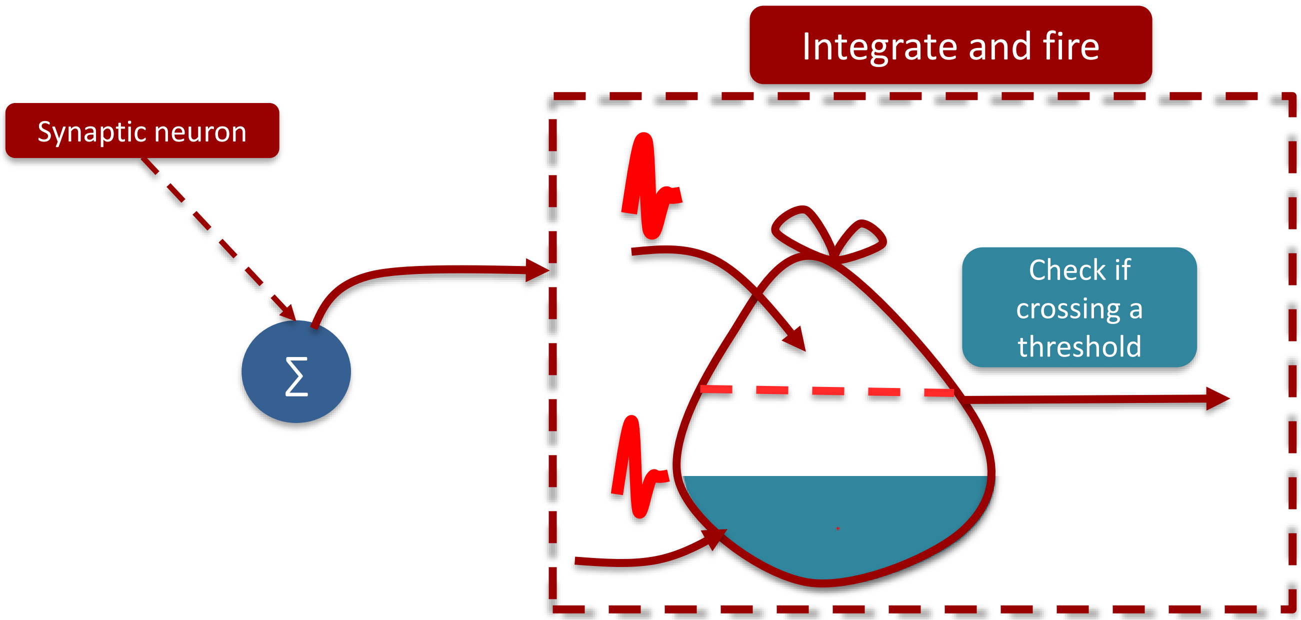
Spike based event-driven compute

Need to look for longer duration to see better

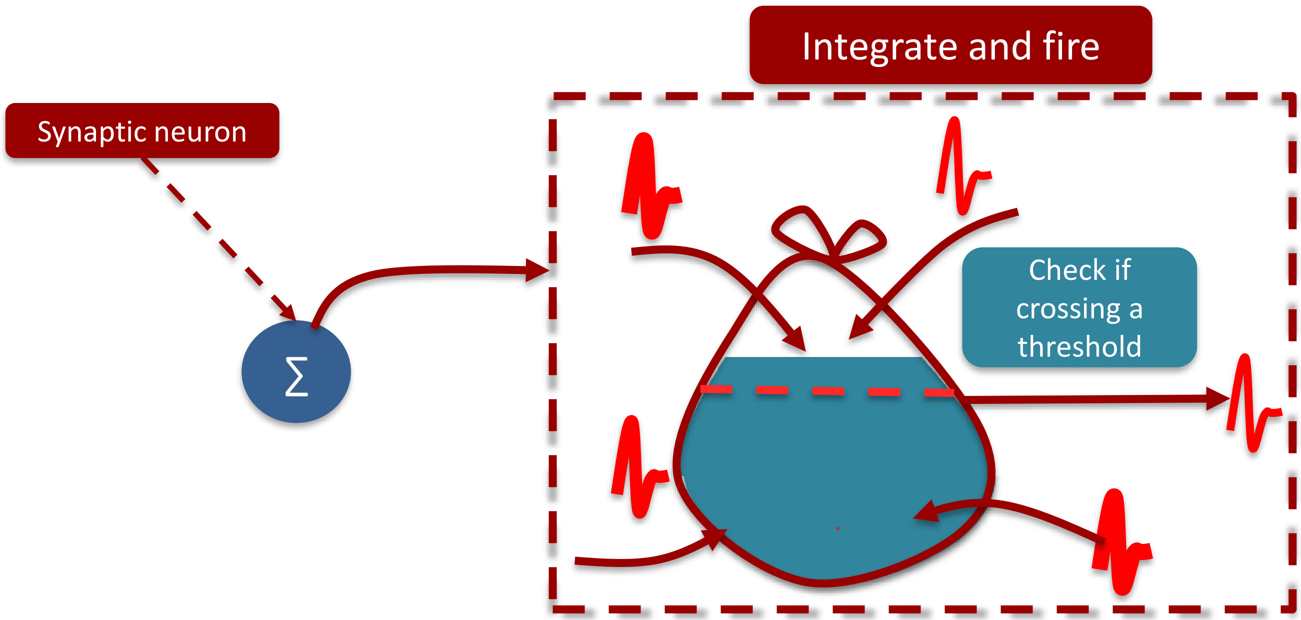
Notion of time

Low-power compute with suitable hardware

Deep SNNs: Beauty and the Beast!

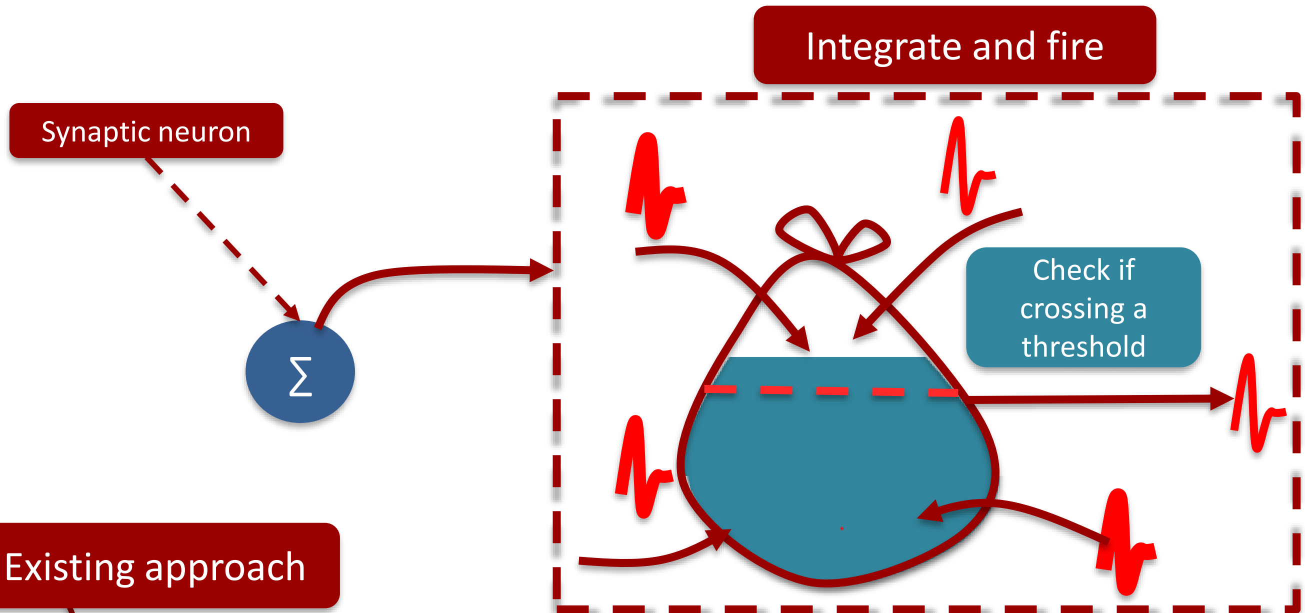


Deep SNNs: Beauty and the Beast!



No Multiply-add!
Only accumulate

Deep SNNs: Beauty and the Beast!



No Multiply-add! Only accumulate

Existing approach

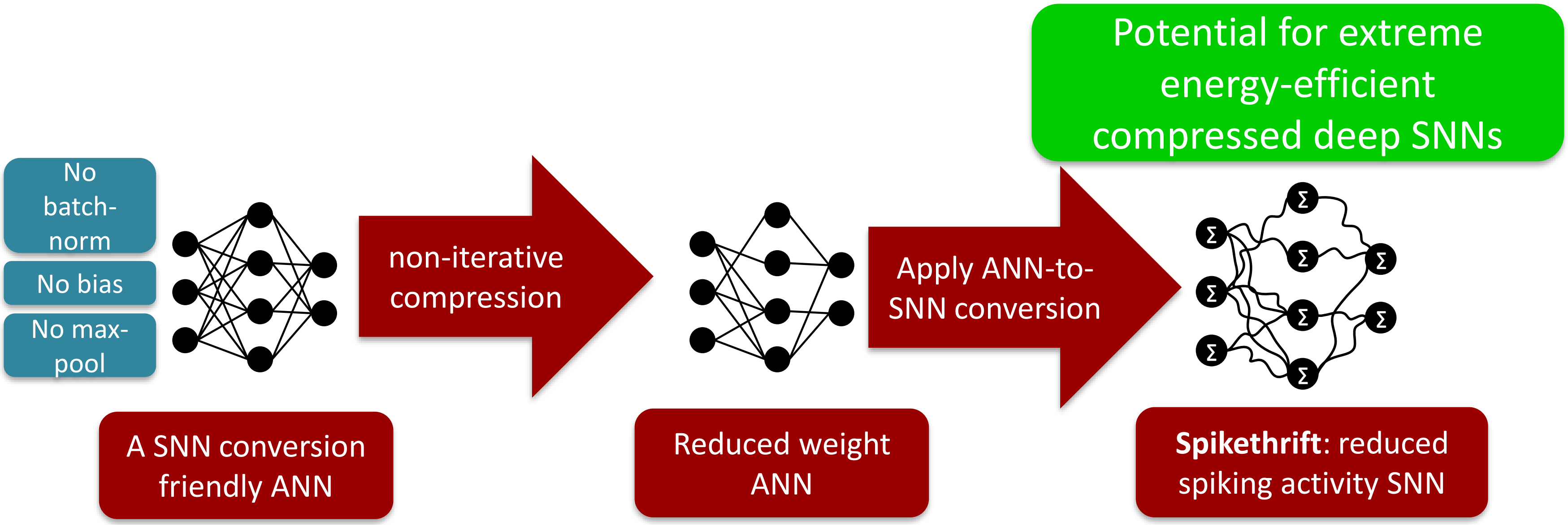
Constrained ANN training → ANN-to-SNN Conversion → SNN inference

- Training time similar to ANN
- Accuracy similar to ANN

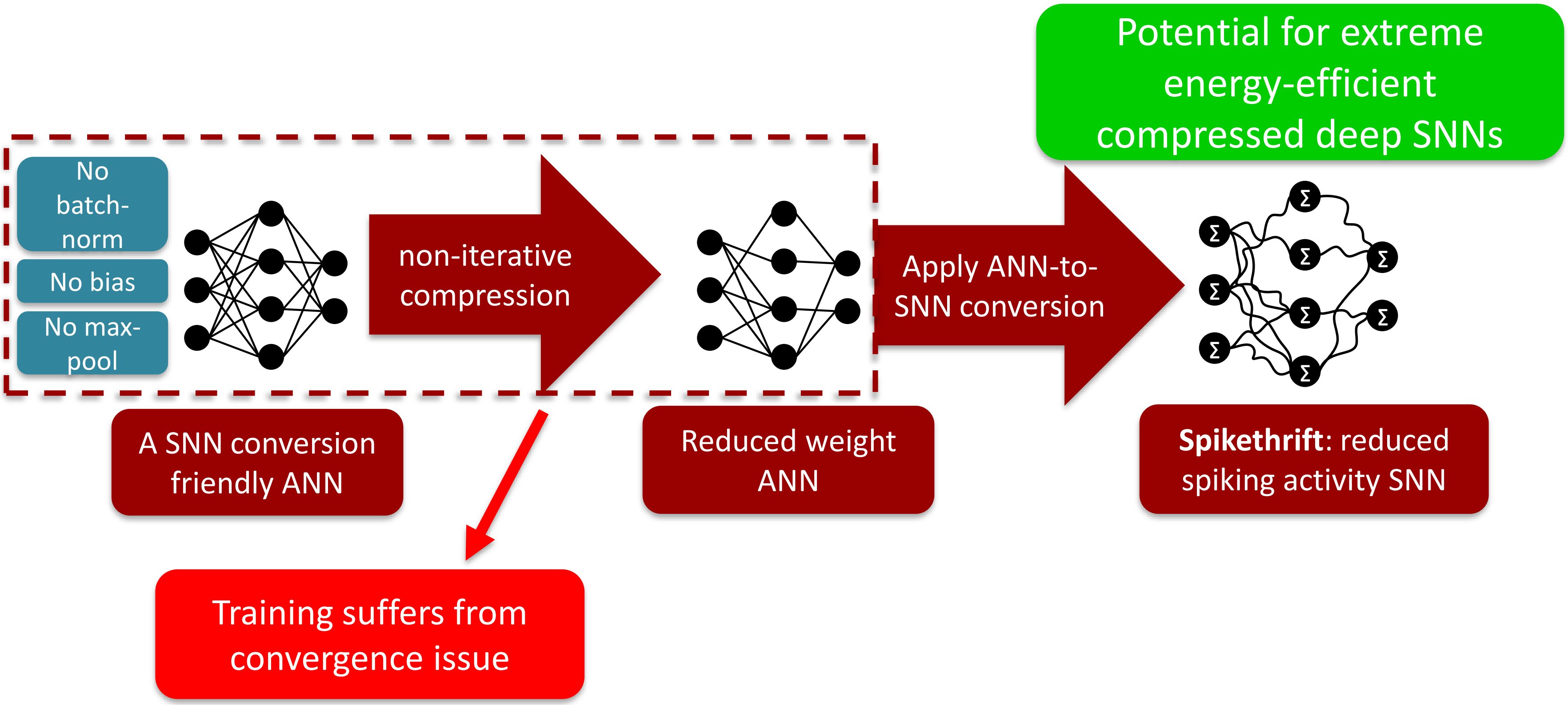
Our focus

- Increased spiking activity over time
- Increased inference energy

Compression for Brain-inspired Computing!

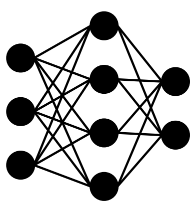


Compression for Brain-inspired Computing!

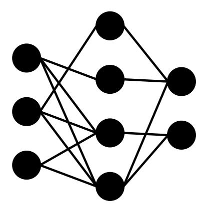


Compression via Brain-inspired Learning!

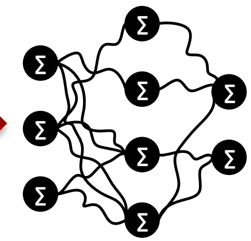
No batch-norm
No bias
No max-pool



non-iterative compression



Apply ANN-to-SNN conversion

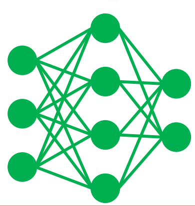


Potential for extreme energy-efficient compressed deep SNNs

A SNN conversion friendly ANN

Reduced weight ANN

Spikethrift: reduced spiking activity SNN



A pre-trained unpruned meta model

Let us use someone to guide this fellow

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

*Based on initial results on CIFAR datasets with VGG model

Summary

Towards a ML driven future: Sustainable and **robust**

- ✔ Reduce training energy through a novel convolution-based model
- ✔ Reduce inference energy and retain robustness through a unified training via a comprehensive loss
- WORK IN PROGRESS: A guided compression strategy for event-driven SNN to yield extreme energy-efficient models

Thanks to all …

QUESTIONS