Spiking Neural Networks: Exploration of Two Key Factors: Sparsity and Robustness

08.21.2021

Souvik Kundu*

Ph.D. Candidate

Electrical and Computer Engineering

USC University of Southern California

*Annenberg Fellow and MHI Scholar Finalist at USC, QIF Finalist.

Virtual USC-IISc Talk

Introduction to the Researcher

Name:
Souvik Kundu.

Hometown:
Kolkata, India.

Current Position:
Ph.D. candidate at University of Southern California.

Concurrent Position:
Research intern at Intel AI Labs, USA.

Past Positions (s):
Design Engineer at Texas Instruments, India.
R & D Engineer at Synopsys, India.

Last Completed Degree:
M. Tech in VLSI, IIT Kharagpur (DR-1).

Web: ksouvik52.github.io

Myself @California, USA, 2020.

Research Timeline: Overview

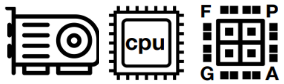
Goal

Energy-efficient DNN training and inference

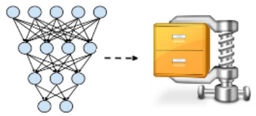
Improved Robustness

Analysis of Privacy

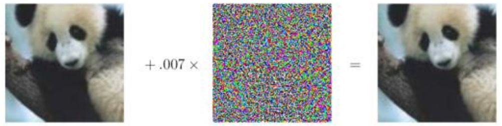
Problem



Pre-defined Sparsity: efficient training



Model compression: efficient inference



Model robustness: safe inference

Model privacy: private inference

2019-2021

Publication

IEEE Trans. on Computers 2020, Allerton 2019, ISVLSI 2019

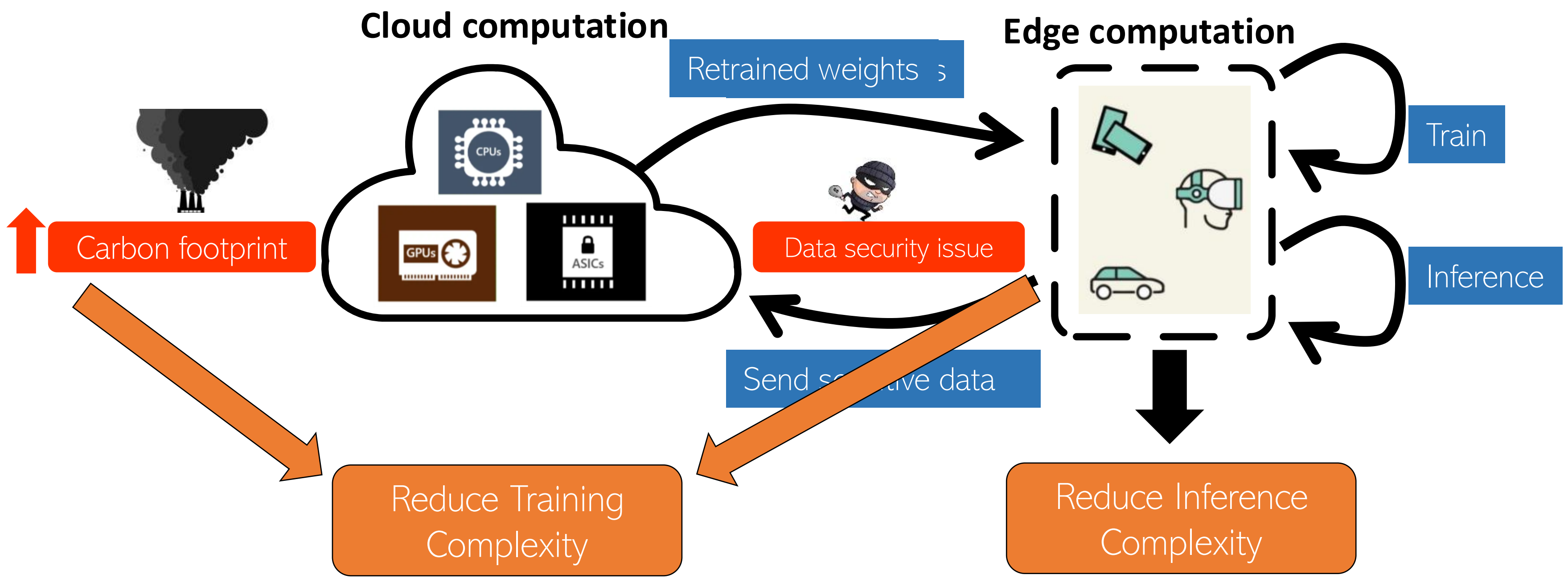
ASP-DAC 2021, WACV 2021, Journal submitted

ASP-DAC 2021, ICCV 2021, Journal submitted

CVPR (W) 2021 NeurIPS 2021 (under review)

Work on Spiking Neural Networks (SNNs)

AI: Energy-Efficiency is a Demand now!



Why Brain-Inspired SNNs?

USC University of Southern California

- Can be extremely compute-energy efficient.
- Can work in an event-driven way on underlying **Neuromorphic** hardware.
- Assumed to mimic functionality of human brain.
- Requires reduced memory for activation storage.

iScience

CellPress REVIEWS

Review

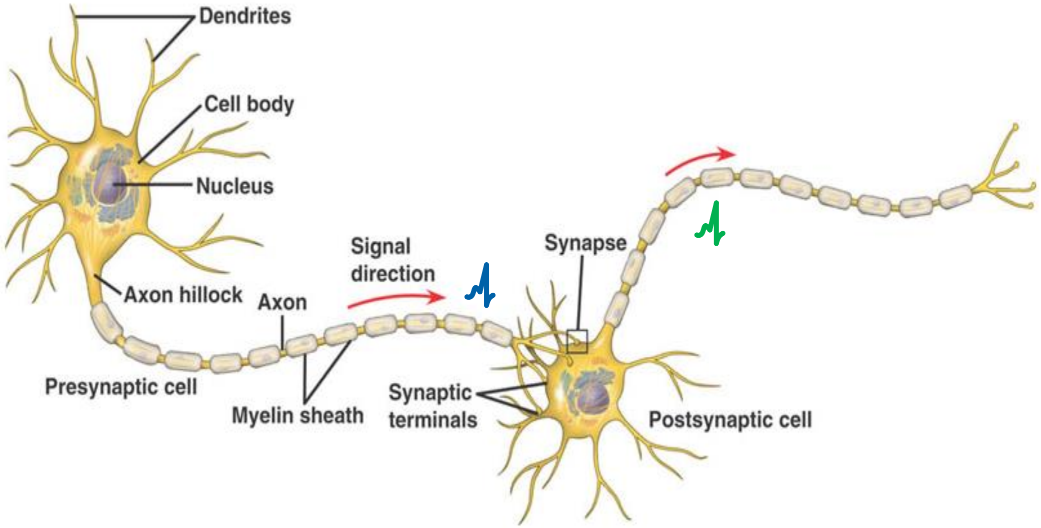
Data and Power Efficient Intelligence with Neuromorphic Learning Machines

Emre O. Neftci[1,2,*]

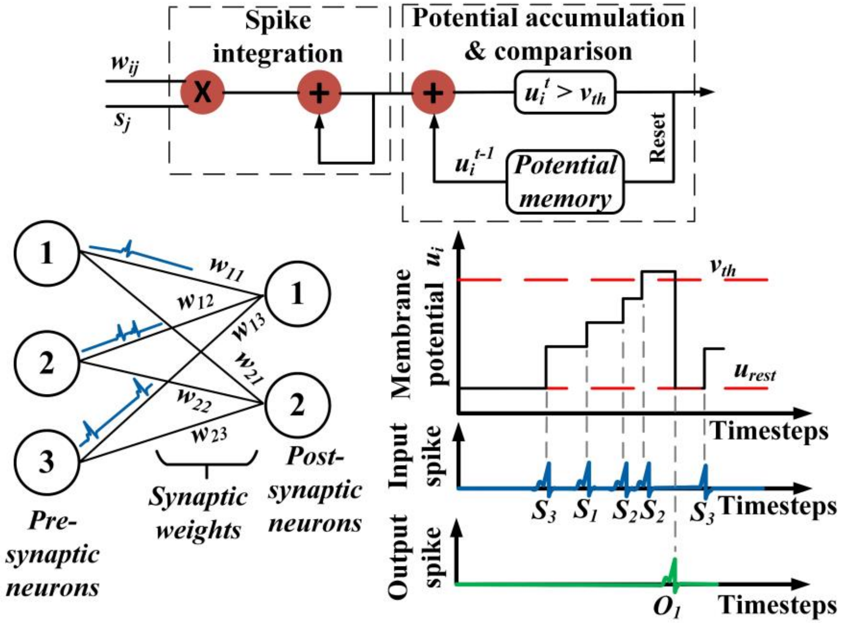
The success of deep networks and recent industry involvement in brain-inspired computing is igniting a widespread interest in neuromorphic hardware that emulates the biological processes of the brain on an electronic substrate. This review explores interdisciplinary approaches anchored in machine learning theory that enable the applicability of neuromorphic technologies to real-world, human-centric tasks. We find that (1) recent work in binary deep networks and approximate gradient descent learning are strikingly compatible with a neuromorphic substrate; (2) where real-time adaptability and autonomy are necessary, neuromorphic technologies can achieve significant advantages over main-stream ones; and (3) challenges in memory technologies, compounded by a tradition of bottom-up approaches in the field, block the road to major breakthroughs. We suggest that a neuromorphic learning framework, tuned specifically for the spatial and temporal constraints of the neuromorphic substrate, will help guiding hardware algorithm co-design and deploying neuromorphic hardware for proactive learning of real-world data.

Image taken from “Data and Power Efficient Intelligence with Neuromorphic Learning Machines”, 2018.

Basics of SNNs



Important components of brain nerve cells



Synaptic weight and event-based Neuromorphic computing

Basic working principle of SNNs

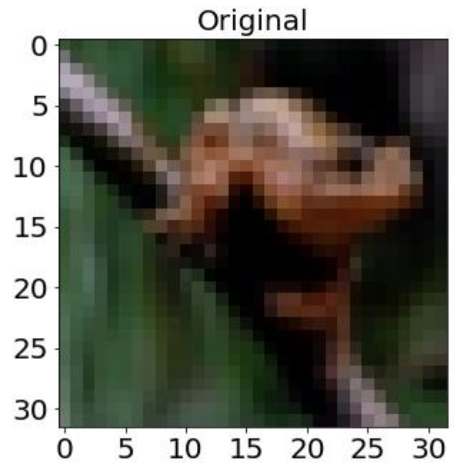
Leaky integrate and fire (LIF) neuron dynamics in discrete time

$$u_i^{t+1} = \lambda u_i^t + \sum_j w_{ij} O_j^t - v_{th} O_i^t$$

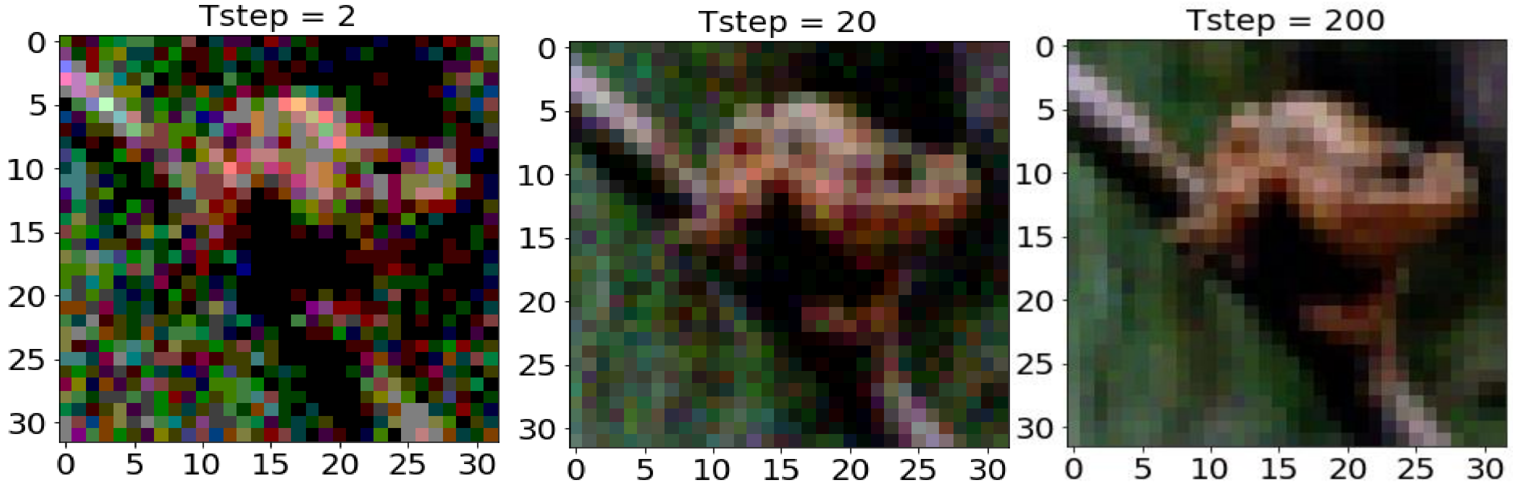
$$O_i^t = \begin{cases} 1, & \text{if } u_i^t > v_{th} \\ 0, & \text{otherwise} \end{cases}$$

- *Corresponds to i^{th} current to one of the pre-synaptic neuron j.*
- *v_{th} - firing threshold voltage of current layer*
- *u_i^{t+1} - potential accumulated at i^{th} neuron at time t+1.*

SNN Training 101



Original image



Rate-coded input for various timesteps

ANN training → ANN-to-SNN conversion → SNN training

Sparsity*

*In this work we term sparse and pruned model interchangeably to mean the same idea of reduced parameter model.

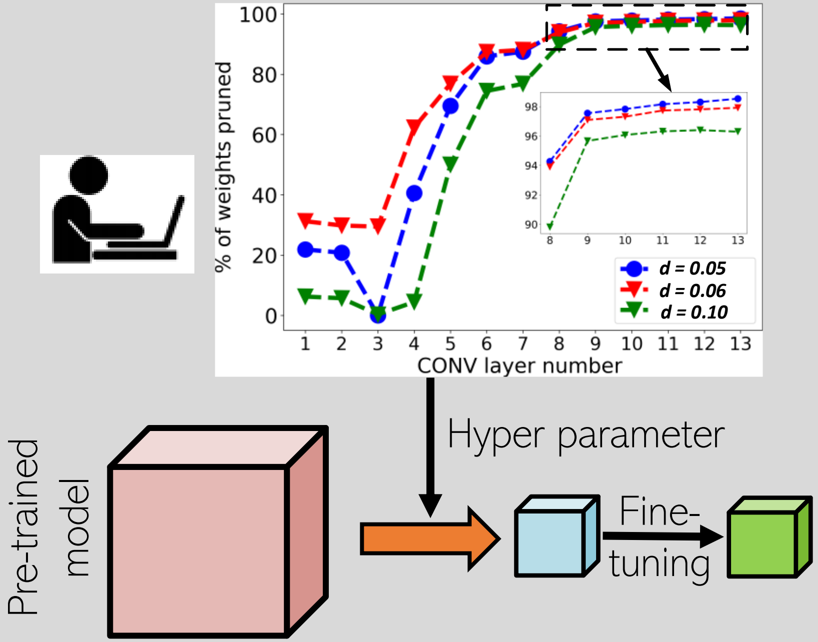
Challenges with Deep SNN models

USC University of Southern California

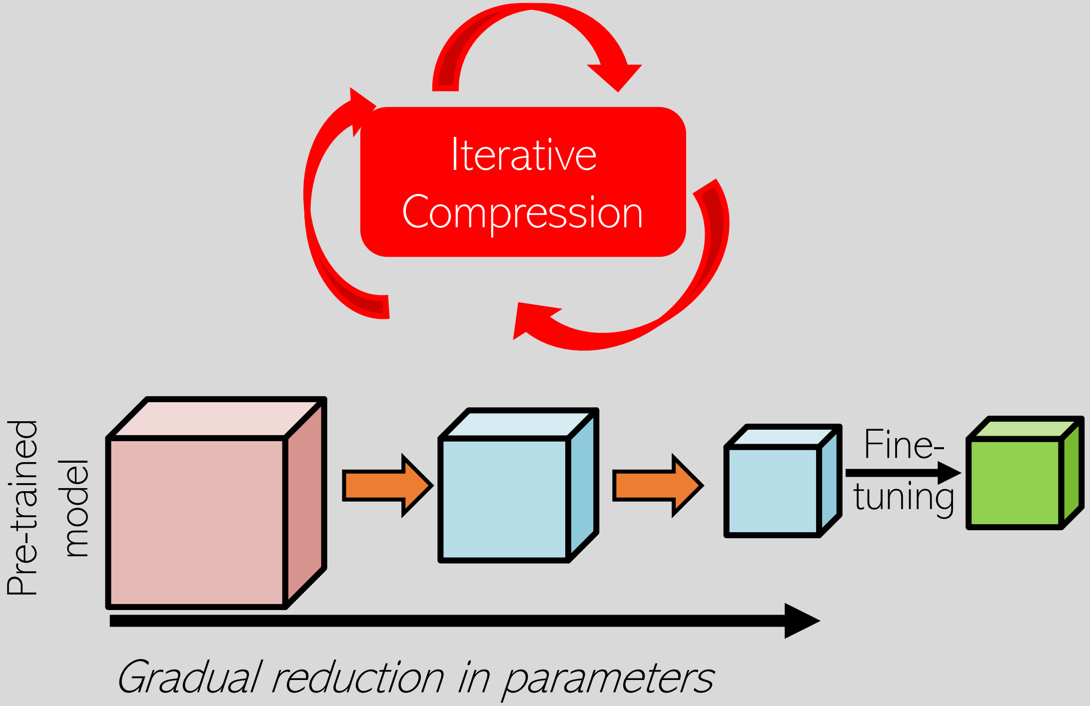
Storage	Similar to ANNs, deep SNNs also suffer from high parameter storage requirement
Convergence	To yield faster ANN-to-SNN convergence SNN models are recommended to not use batch-normalization (BN) layers
Training time	Due to back-prop through time (BPTT) SNNs require orders of larger training time, thus iterative pruning is difficult.

Currently Existing SNN Pruning Schemes

Alternating Direction Method of Multiplier (ADMM)



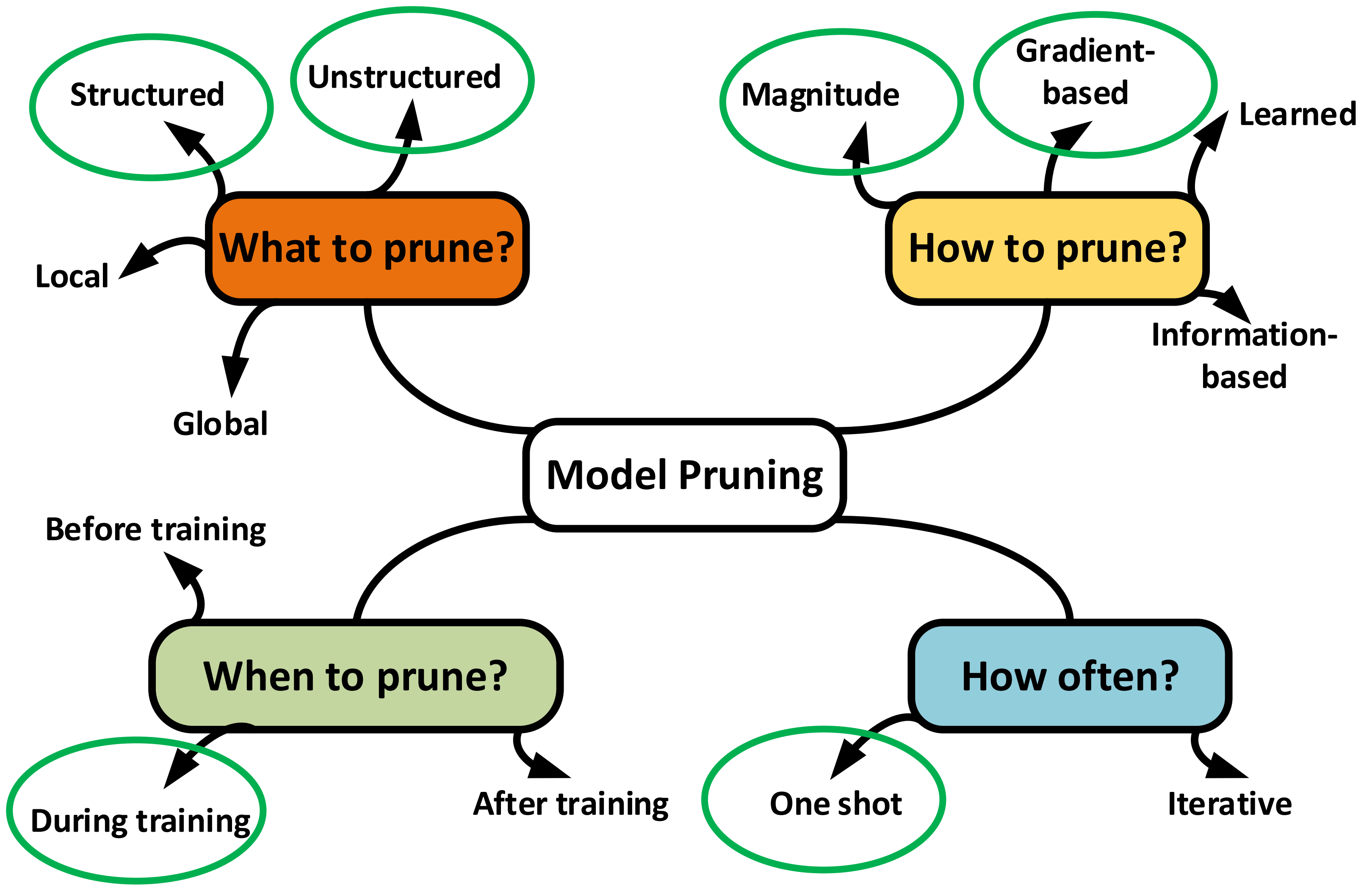
Iterative Magnitude Pruning (IMP)



Increased SNN training time

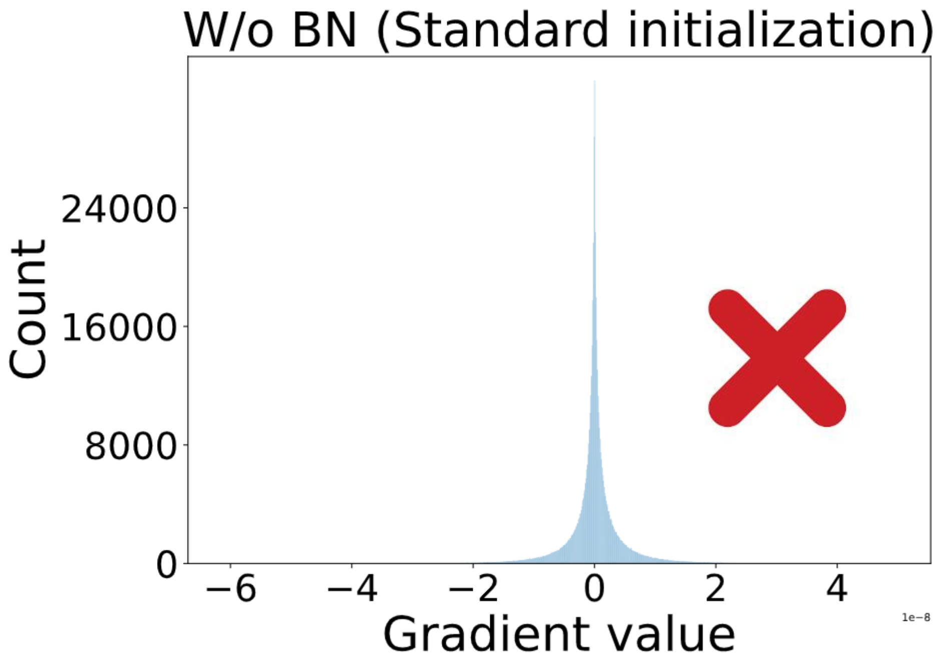
Poor compression ratio

What We Plan to Achieve?

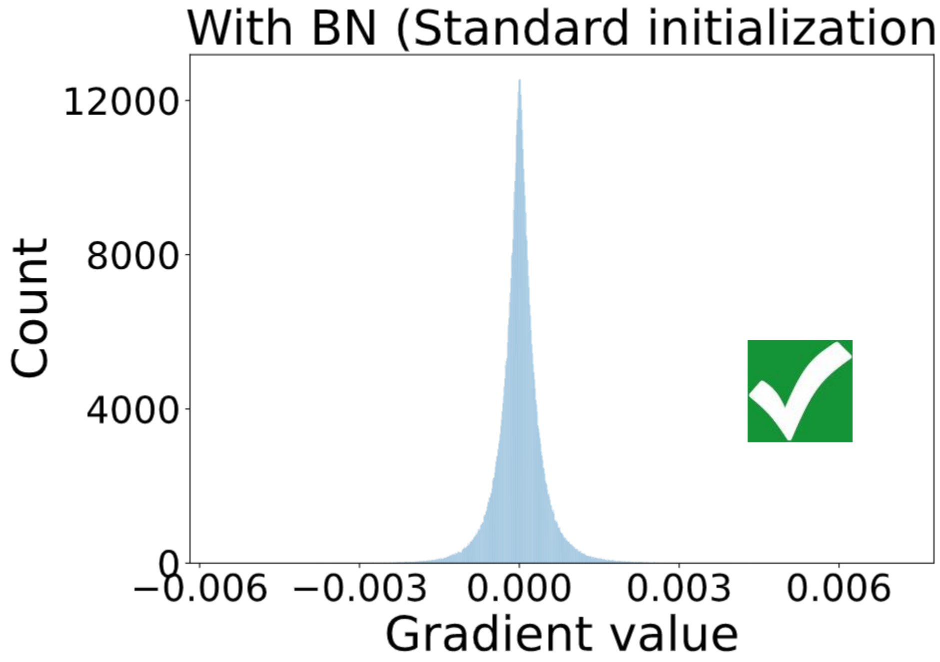


The Problem

Convergence issue

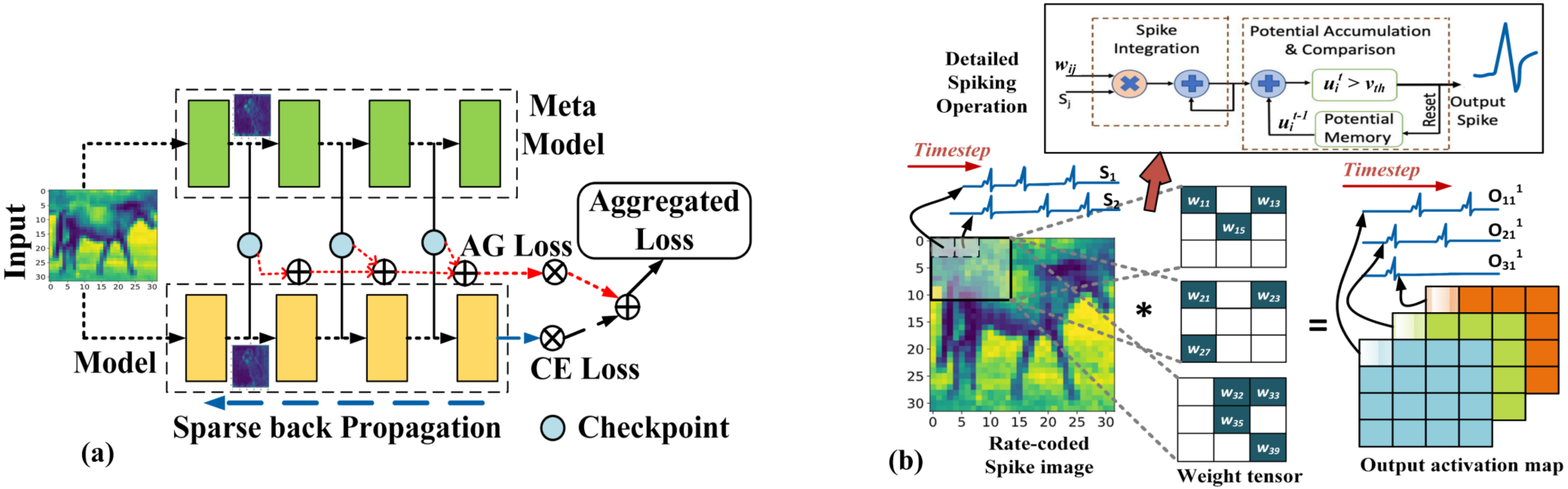


No Convergence issue



Simple sparse learning approaches like DNR fails to compress

Proposed: Attention-Guided Compression (AGC)



Step (a)

ANN training With AGC

ANN-to-SNN conversion

Step (b)

SNN training with sparse-learning

- S. Kundu et al., "Spike-Thrift: Towards Energy-Efficient Deep Spiking Neural Networks by Limiting Spiking Activity via Attention-Guided Compression", WACV, 2021.

AGC-Proposed Training Loss

USC University of Southern California

ANN training

$$\mathcal{L} = \underbrace{\frac{\alpha}{2}\sum_{j\in\mathcal{I}}\left\| \frac{Q_j^{\Psi_c}}{\|Q_j^{\Psi_c}\|_2} - \frac{Q_j^{\Psi_m}}{\|Q_j^{\Psi_m}\|_2}\right\|_2}_{\text{AT loss}} + \underbrace{\mathcal{L}_{CE}^{\Psi_c}(y,\tilde{y})}_{\text{CE loss}}$$

$$w_{ij} = w_{ij} - m_{ij} * \eta\delta w_{ij}$$

Dynamic mask

SNN training

Static mask

$$u_i^{t+1} = u_i^t + \sum_j m_{ij} * w_{ij} O_j^t - v_{th} O_i^t$$

$$O_i^t = \begin{cases} 1, & \text{if } z_i^t > 0, \\ 0, & \text{otherwise} \end{cases}$$

Membrane potential update

$$w_{ij} = w_{ij} - \eta\delta w_{ij}$$

$$\delta w_{ij} = m_{ij} * \sum_t \frac{\partial \mathcal{L}}{\partial O_i^t}\frac{\partial O_i^t}{\partial z_i^t}\frac{\partial z_i^t}{\partial u_i^t}\frac{\partial u_i^t}{\partial w_{ij}^t}$$

Weight update

$$\frac{\partial O_i^t}{\partial z_i^t} = \gamma * max\{0, 1 - |z_i^t|\}$$

Linear Surrogate Gradient

- S. Kundu et al., "Spike-Thrift: Towards Energy-Efficient Deep Spiking Neural Networks by Limiting Spiking Activity via Attention-Guided Compression", *WACV, 2021*.

Results

Authors	Training type	Architecture	Compression ratio	Accuracy (%)	Time steps
Dataset : CIFAR-10					
Cao et al. (2015) [4]	ANN-SNN conversion	3 CONV, 2 linear	1×	77.43	400
Sengupta et al. (2019)[37]	ANN-SNN conversion	VGG16	1×	91.55	2500
Wu et al. (2019) [44]	Surrogate gradient	5 CONV, 2 linear	1×	90.53	12
Rathi et al. (2020) [36]	Hybrid training	VGG16	1× 1×	91.13 **92.02**	100 200
Deng et al. (2020) [8]	STBP training	11 layer CNN	1×	89.53	**8**
Deng et al. (2020) [8]	STBP training	11 layer CNN	4×	87.38	8
This work	Hybrid SL	VGG16	2.5× **33.4**×	91.29 90.15	100 100
Dataset : CIFAR-100					
Deng et al. (2020) [8]	STBP training	11 layer CNN	2×	57.83	8
This work	Hybrid SL	VGG11	**4**×	**64.98**	120

Better accuracy vs. compression ratio trade-off

Table 2. Performance comparison of the proposed hybrid SL with state-of-the-art deep SNNs on CIFAR-10 and CIFAR-100.

- S. Kundu et al., "Spike-Thrift: Towards Energy-Efficient Deep Spiking Neural Networks by Limiting Spiking Activity via Attention-Guided Compression", *WACV, 2021*.

Summary

- Proposed AGC can yield compressed SNN models through a one-shot pruning of the target model.
- AGC achieves SOTA compressed model that can retain classification performance.
- AGC finds optimal layer significance for a given target global pruning ratio-no need of manual or search or separate learning techniques to evaluate layer significance.

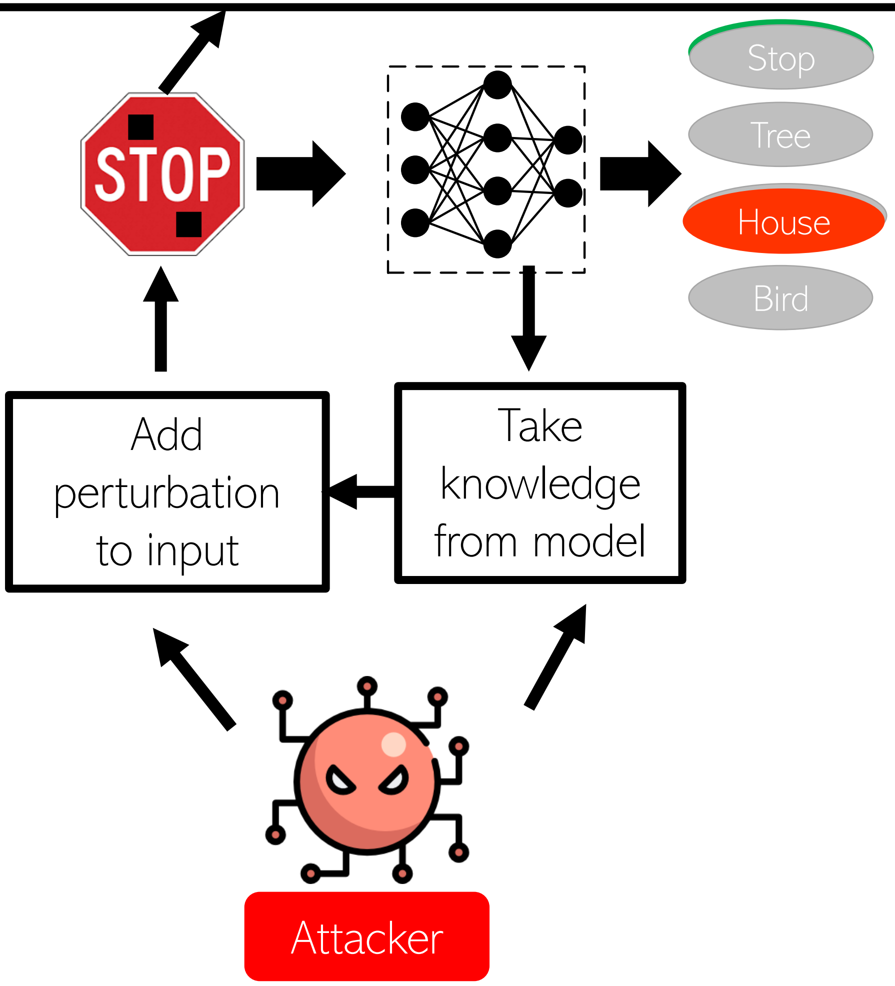
Fundamental take-away:

Exploding gradient issue of BN-less models can be resolved through guidance via activation maps from a trained model

Robustness

Robustness is a Growing Concern

Robustness is model performance against the perturbed inputs



A life-threatening consequence

How are the Perturbations Generated?

USC University of Southern California

Gradient-based strong perturbation generation

- Fast gradient sign method (FGSM) → $\hat{\boldsymbol{x}} = \boldsymbol{x} + \epsilon \times sgn(\nabla_x J(g(\boldsymbol{x}; \boldsymbol{\theta}), \boldsymbol{t}))$
- Projected gradient descent method (PGD) → $\hat{\boldsymbol{x}}^k = \mathrm{Proj}_{P_\epsilon(\boldsymbol{x})}(\hat{\boldsymbol{x}}^{k-1} + \alpha \times sgn(\nabla_x J(g(\hat{\boldsymbol{x}}^{k-1}; \boldsymbol{\theta}), \boldsymbol{t})))$

Are SNNs Inherently Robust Against Adversary?

USC University of Southern California

Inherent Adversarial Robustness of Deep Spiking Neural Networks: Effects of Discrete Input Encoding and Non-Linear Activations

Saima Sharmin[1][0000−0002−1866−9138], Nitin Rathi[1][0000−0003−0597−064X], Priyadarshini Panda[2][0000−0002−4167−6782], and Kaushik Roy[1][0000−0002−0735−9695]

[1] Purdue University, West Lafayette IN 47907, USA
{ssharmin,rathi2,kaushik}@purdue.edu
[2] Yale University, New Haven CT 06520, USA
priya.panda@yale.edu

ECCV 2020.

Securing Deep Spiking Neural Networks against Adversarial Attacks through Inherent Structural Parameters

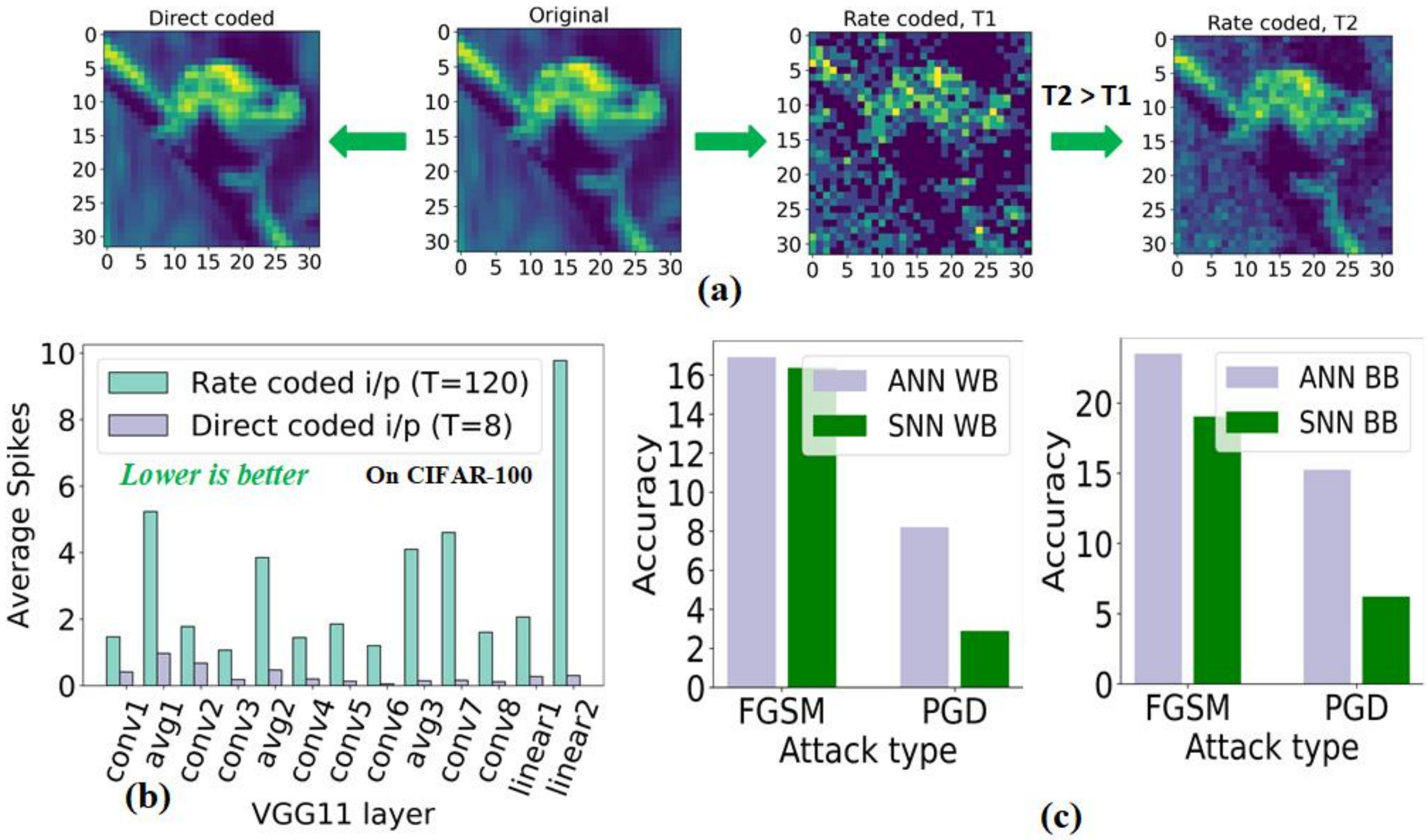
Rida El-Allami[1,*], Alberto Marchisio[2,*], Muhammad Shafique[3], Ihsen Alouani[1]
[1] *IEMN CNRS-UMR8520, Université Polytechnique Hauts-De-France, Valenciennes, France*
[2] *Institute of Computer Engineering, Technische Universität Wien, Vienna, Austria*
[3] *Division of Engineering, New York University Abu Dhabi, UAE*
Email: rida.elallami@etu.uphf.fr, alberto.marchisio@tuwien.ac.at, muhammad.shafique@nyu.edu, ihsen.alouani@uphf.fr

DATE 2021.

- Few earlier research have concluded that SNNs **are to some extent**, inherently robust to adversarial images.
- Earlier research also hinted at SNNs to be **more inherently robust** than ANN counter-parts.
- However, **no earlier work** has concluded the same for extremely low-latency SNNs, which is a **more applicable** scenario for real-time applications.

The Problem

- Low-latency direct input SNNs (LLSNNs) are extremely compute-efficient.
- However, these SNNs sacrifice adversarial robustness significantly.
- Low-latency SNNs has poor adversarial robustness compared to ANN counter-parts.



S. Kundu et al., "HIRE-SNN: Harnessing the Inherent Robustness of Energy-Efficient Deep Spiking Neural Networks by Training with Crafted Input Noise", *ICCV 2021*.

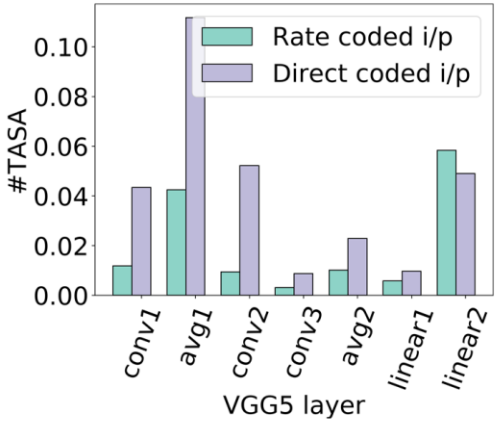
Where do LLSNNs Differ from Rate-Coded SNNs?

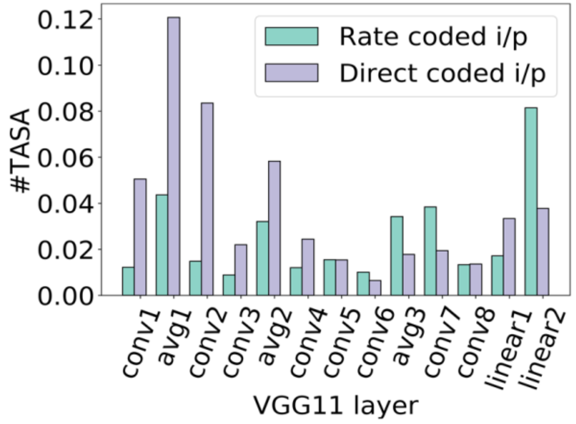
USC University of Southern California

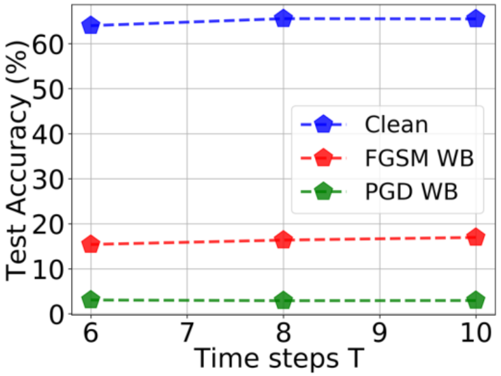
- ❖ Activation-sparsity is helpful for robustness: Spiking-activity per unit time step is more in LLSNNs

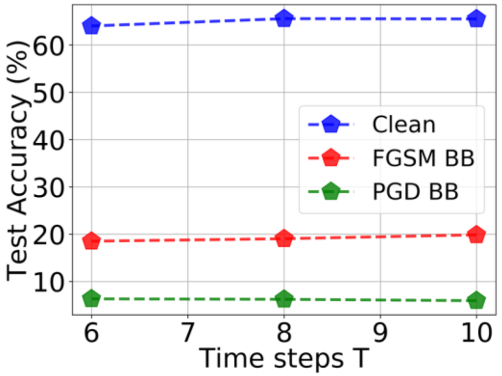
- ❖ Input approximation is helpful for robustness: Direct input makes sure no input approximation happens

- ❖ Reduction in time-step helps improve robustness: However, LLSNNs can't gain from further reduction in t-steps.





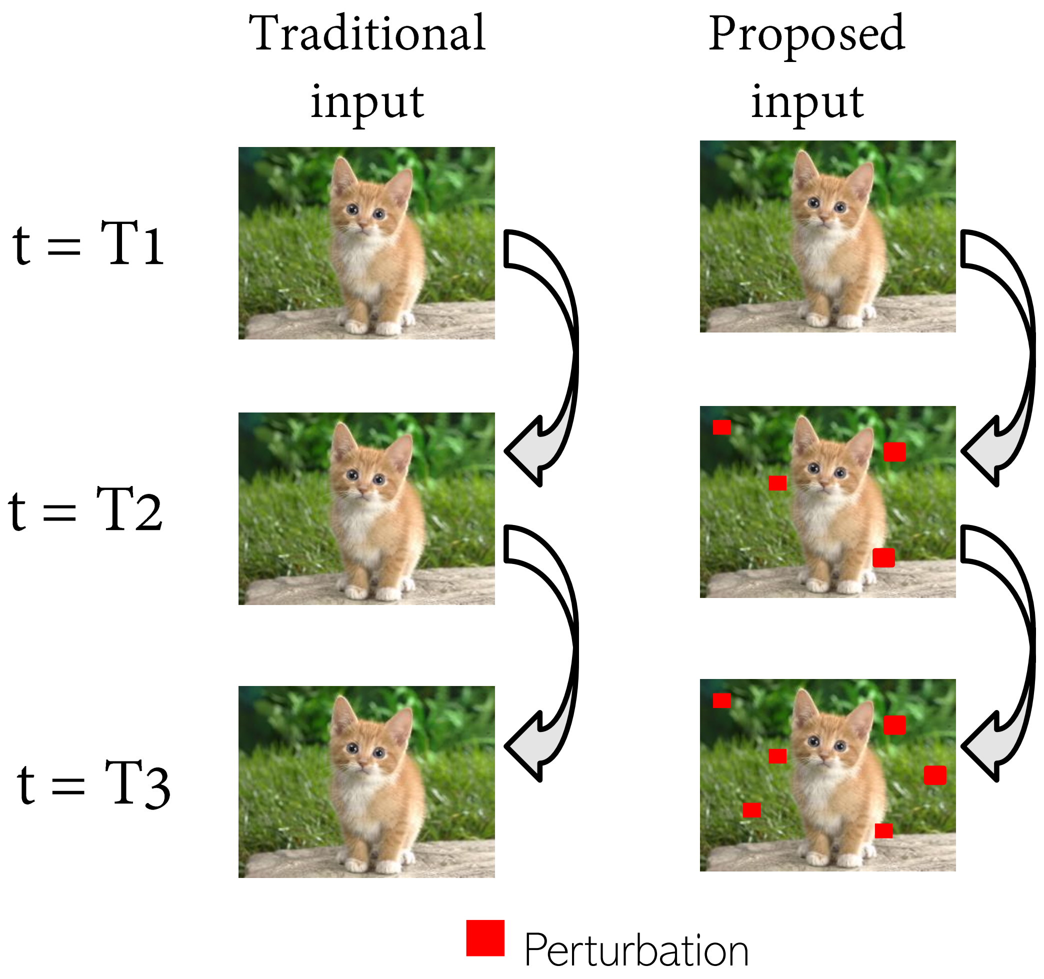




S. Kundu et al., "HIRE-SNN: Harnessing the Inherent Robustness of Energy-Efficient Deep Spiking Neural Networks by Training with Crafted Input Noise", *ICCV 2021.*

Proposed Training Scheme: HIRE-SNN

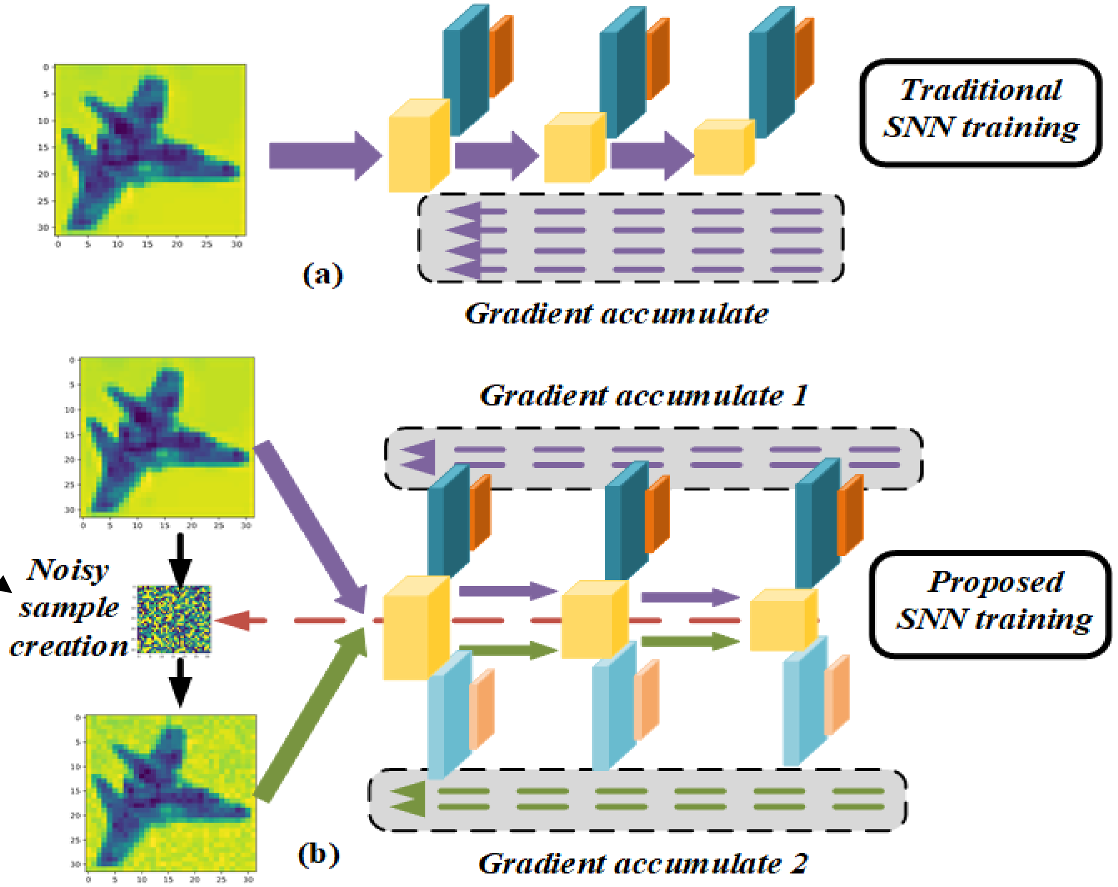
- Partitioning the t-steps T into multiple periods of small steps.
- Instead of using the same image over multiple steps, feed different perturbed variants of the image, during different periods.



S. Kundu et al., "HIRE-SNN: Harnessing the Inherent Robustness of Energy-Efficient Deep Spiking Neural Networks by Training with Crafted Input Noise", *ICCV 2021*.

HIRE-SNN: Training Strategy

$$\kappa = clip[\kappa + \epsilon_s \times sign(\nabla_x \mathcal{L}), -\epsilon_t, +\epsilon_t]$$



S. Kundu et al., "HIRE-SNN: Harnessing the Inherent Robustness of Energy-Efficient Deep Spiking Neural Networks by Training with Crafted Input Noise", *ICCV 2021*.

HIRE-SNN: Performance

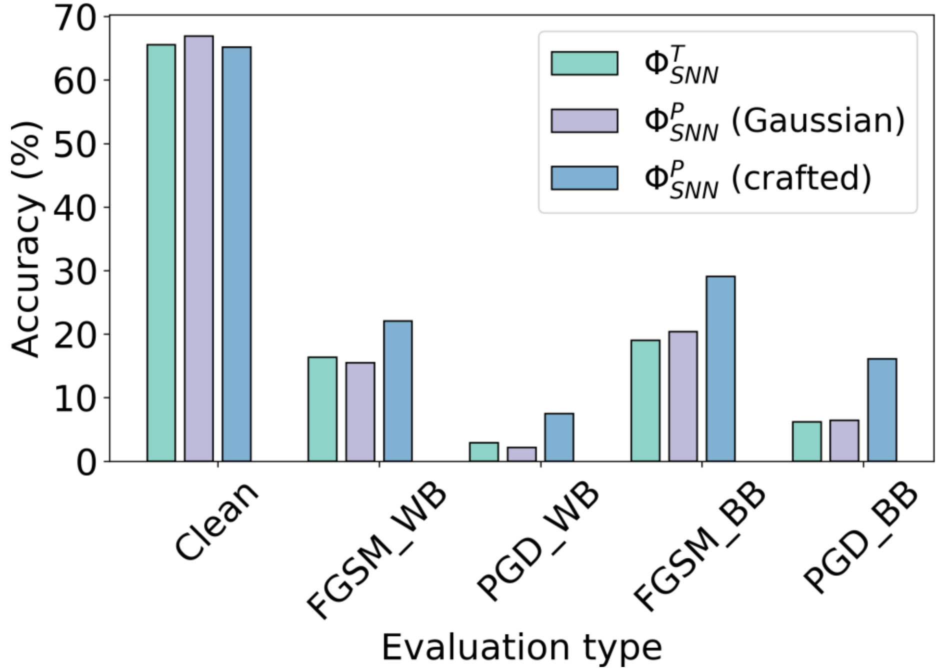
Model	Accuracy (%) with proposed SNN training			Δ_a over traditional SNN training		Δ_a over ANN equivalent	
	Clean(Δ_d)	FGSM	PGD	FGSM	PGD	FGSM	PGD
Dataset : CIFAR-10							
VGG5	87.5 (-0.4)	38.0	9.1	+2.5	**+3.8**	**+25**	**+7.1**
ResNet12	90.3 (-1.6)	33.3	3.8	**+12.2**	+3.5	+13.4	+1.8
Dataset : CIFAR-100							
VGG11	65.1 (-0.4)	22.0	7.5	+5.7	+4.6	+5.1	-0.7
ResNet12	58.9 (-3.0)	19.3	5.3	**+8.8**	**+4.7**	**+5.8**	**+2.5**

Model	Accuracy (%) with proposed SNN training			Δ_a over traditional SNN training		Δ_a over ANN equivalent	
	Clean	FGSM	PGD	FGSM	PGD	FGSM	PGD
Dataset : CIFAR-10							
VGG5	87.5	42.1	14.9	+3.9	**+8.3**	**+18.1**	**+8.5**
ResNet12	90.3	38.4	7.8	**+13.7**	+7.2	+9.7	+3.5
Dataset : CIFAR-100							
VGG11	65.1	29.1	16.1	+10.0	+9.9	**+5.6**	**+0.9**
ResNet12	58.9	24.5	12.1	**+10.4**	**+10.1**	+1.3	$\sim$0

HIRE-SNN consistently outperforms, traditional SNNs in providing better robustness

S. Kundu et al., "HIRE-SNN: Harnessing the Inherent Robustness of Energy-Efficient Deep Spiking Neural Networks by Training with Crafted Input Noise", *ICCV 2021*.

Crafted Noise vs. Gaussian Noise



Gaussian noise induced inputs does not improve performance against strong adversary

S. Kundu et al., "HIRE-SNN: Harnessing the Inherent Robustness of Energy-Efficient Deep Spiking Neural Networks by Training with Crafted Input Noise", *ICCV 2021*.

Summary

- Inherent robustness of LLSNNs (direct input) are poorer compared to rate-coded SNNs, when trained in traditional approach.

- HIRE-SNNs is a novel training strategy that can train SNNs with improved robustness against adversary.

- Crafted input noise helps improve robustness, however simple noise addition (e.g.: Gaussian noise) doesn't help against strong adversary.

Fundamental take-away:
A fixed image over the whole window of t-steps is not necessary for the SNN to train, various augmented variants can be fed to improve performance.

Recent Publications

USC University of Southern California

1. [ICCV 2021] *S. Kundu* et al., "HIRE-SNN: Harnessing the Inherent Robustness of Energy-Efficient Deep Spiking Neural Networks by Training with Crafted Input Noise".
2. [CVPRW 2021] *S. Kundu* et al., "Skeptical Student: Diminishing the Effect of Leaking Teacher in Knowledge Distillation".
3. [ICASSP 2021] *S. Kundu* et al., "AttentionLite: Towards Efficient Self-Attention Models for Vision".
4. [WACV 2021] *S. Kundu* et al., "Spike-Thrift: Towards Energy-Efficient Deep Spiking Neural Networks by Limiting Spiking Activity via Attention-Guided Compression".
5. [ASP-DAC 2021] *S. Kundu* et al., "DNR: A Tunable Robust Pruning Framework Through Dynamic Network Rewiring of DNNs".
6. [IEEE TC submit] *S. Kundu* et al., "Towards Low-Latency Energy-Efficient Deep SNNs via Attention-Guided Compression".
7. [IEEE TC 2020] *S. Kundu* et al., "Pre-defined Sparsity for Low-Complexity Convolutional Neural Networks".
8. [IJCNN 2021] *G. Datta, S. Kundu,* et al., "Training Energy-Efficient Deep Spiking Neural Networks with Single-Spike Hybrid Input Encoding"
9. [ACM TECS submit] *S. Kundu* et al., "Towards Adversary aware Non-Iterative Model Pruning Through Dynamic Network Rewiring of DNNs".

[N.B.: For full list please visit: ksouvik52.github.io]

Thank You!

Questions