On the Analysis of Model Robustness and Privacy in Various Compression Frameworks

By

Souvik Kundu

Ph.D. Candidate

Ming Hsieh Department of Electrical and Computer Engineering

USC University of Southern California

A Brief Introduction about Me

USC University of Southern California

Birthplace: Kolkata, India

Latest completed degree: M.Tech.

University: IIT Kharagpur, India

Prior work experiences: Texas Instruments, India; Synopsys, India

Current position: started 5th year of Ph.D. at USC

Concurrent position: Research intern, Intel AI Labs, USA

Current research focus: Energy-efficiency, robustness, and privacy in A.I.

Webpage: ksouvik52.github.io

Advisors:

Dr. Massoud Pedram

Dr. Peter A. Beerel

Outline of the Talk

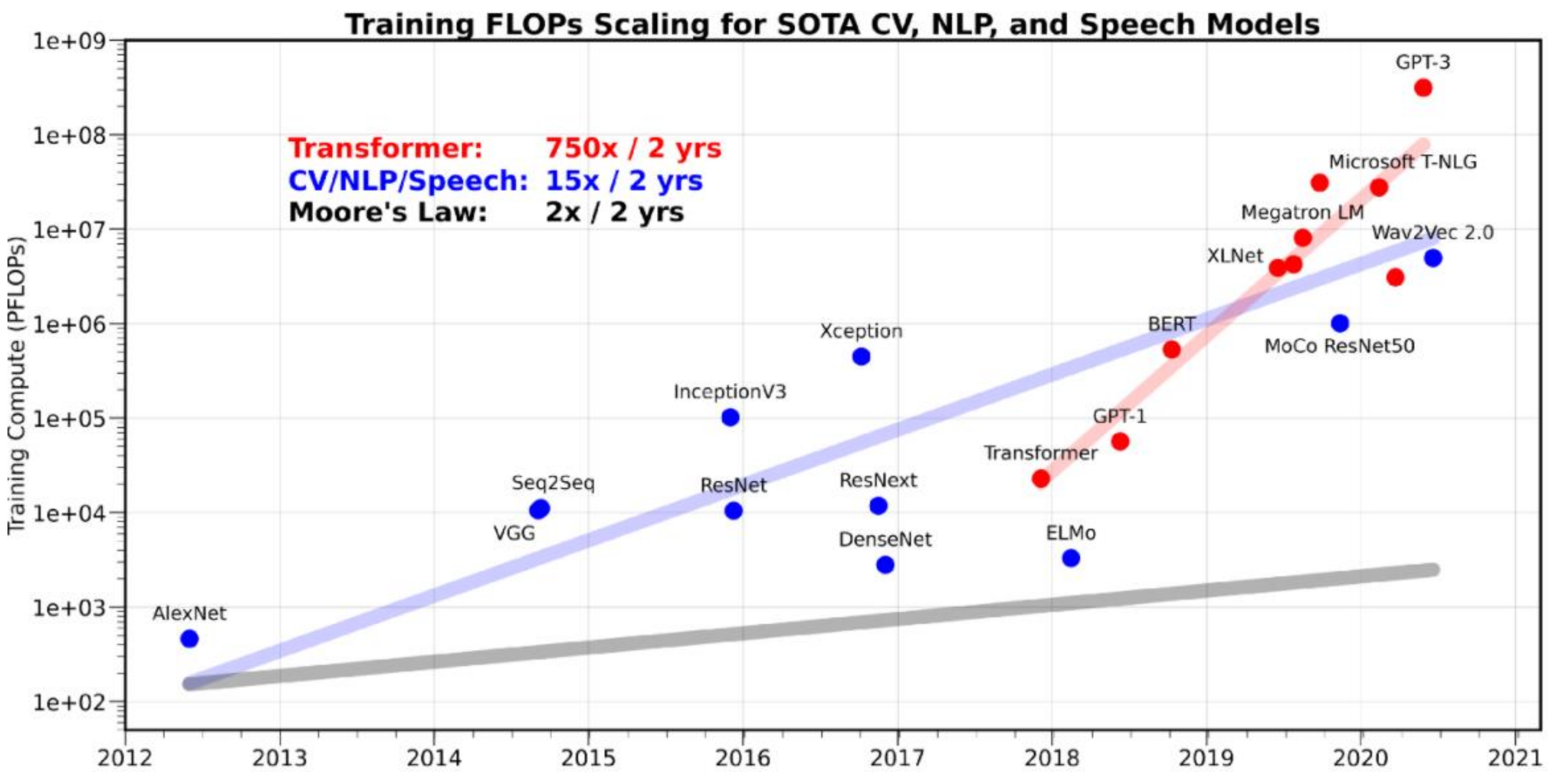
USC University of Southern California

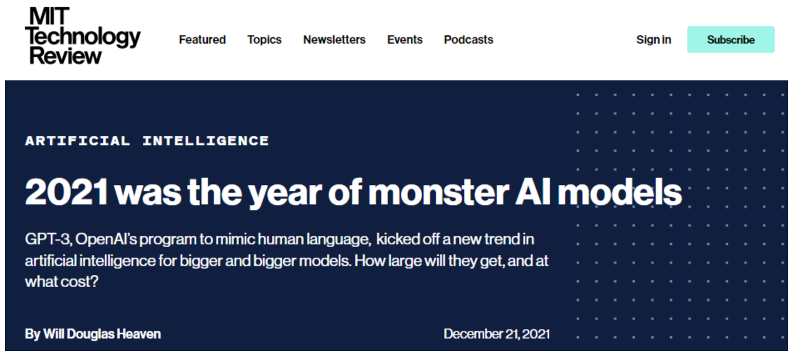
- ➢ Robustness in model pruning framework.
- ➢ Robustness for brain-inspired models.
- ➢ Model privacy in distillation framework.
- ➢ Future research discussion and conclusion.

1a: Robustness for Pruned Models

Growing Concern of A.I. and Memory Wall Problem

USC University of Southern California




MIT Technology Review

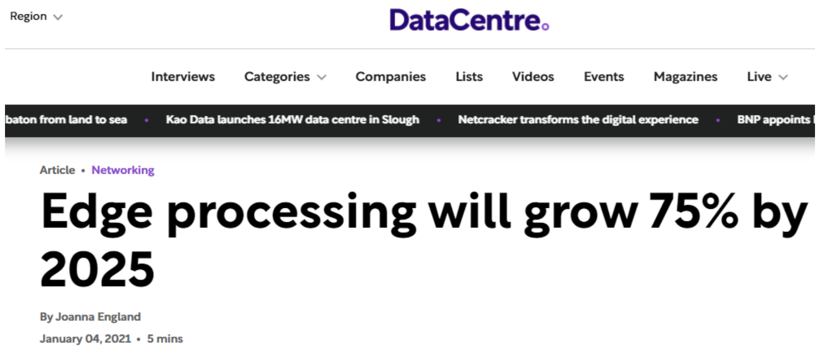
Featured Topics Newsletters Events Podcasts Sign in Subscribe

ARTIFICIAL INTELLIGENCE

2021 was the year of monster AI models

GPT-3, OpenAI's program to mimic human language, kicked off a new trend in artificial intelligence for bigger and bigger models. How large will they get, and at what cost?

By Will Douglas Heaven December 21, 2021


Region

DataCentre.

Interviews Categories Companies Lists Videos Events Magazines Live

baton from land to sea • Kao Data launches 16MW data centre in Slough • Netcracker transforms the digital experience • BNP appoints

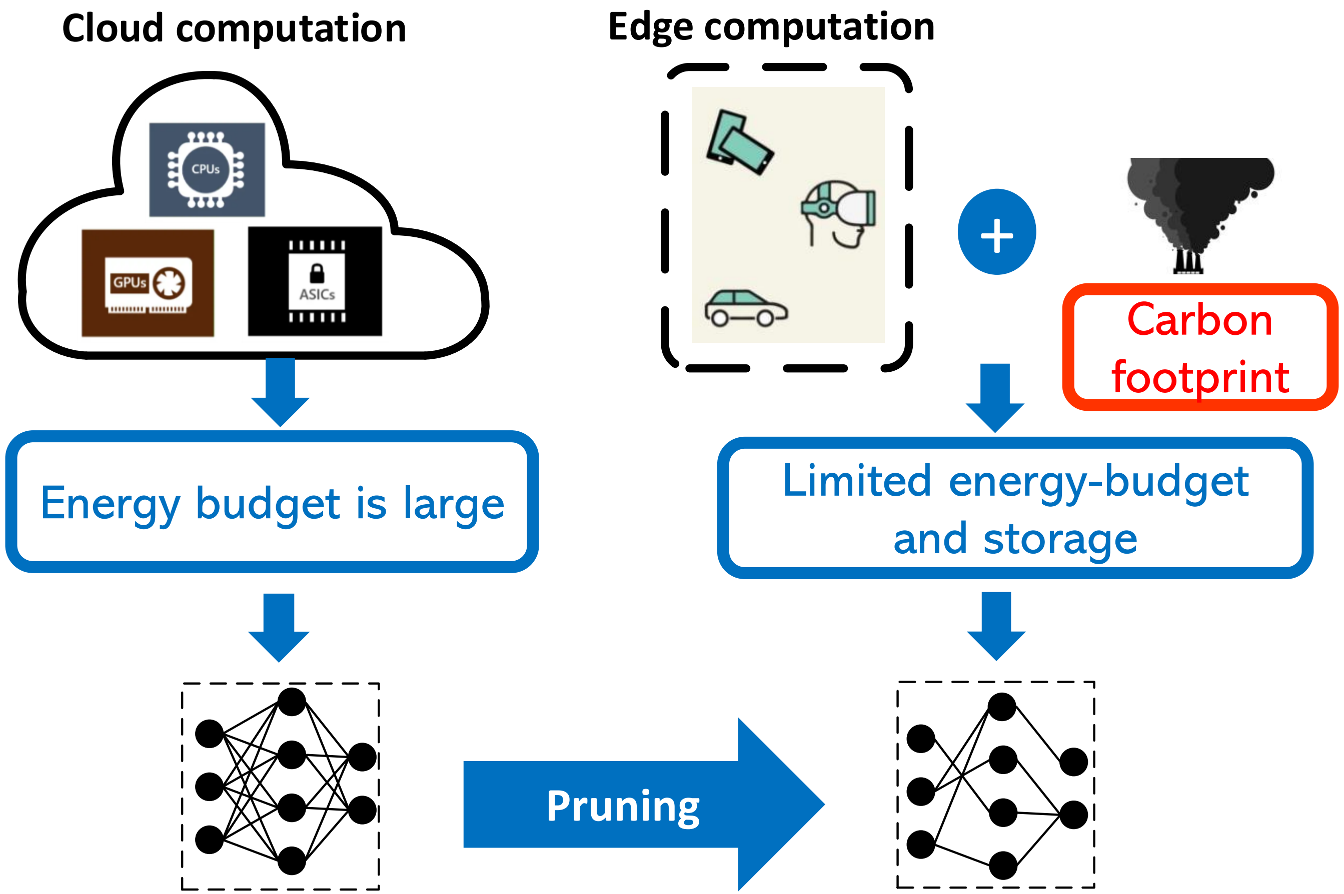
Article • Networking

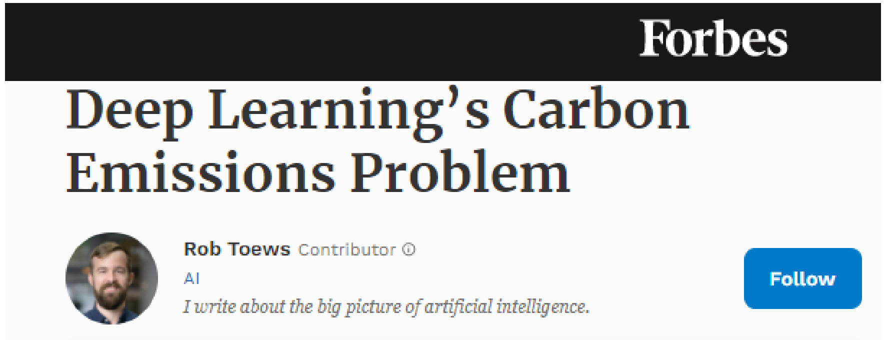
Edge processing will grow 75% by 2025

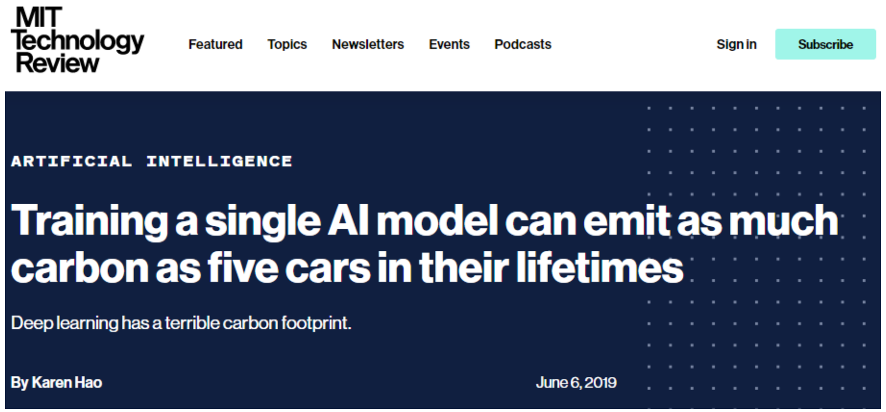
By Joanna England
January 04, 2021 • 5 mins

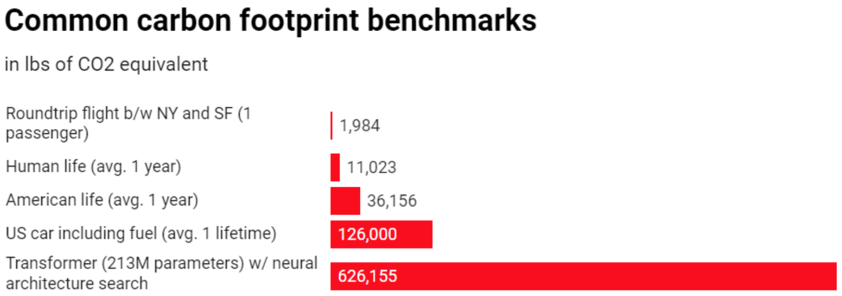
Plot courtesy: https://medium.com/riselab/ai-and-memory-wall-2cb4265cb0b8

Model Pruning is Necessary



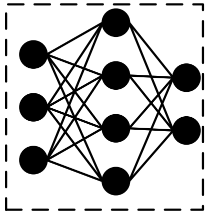






Model Robustness is Necessary as Well

USC University of Southern California



Stop

Tree

House

Bird

Add perturbation to input

kn
from model

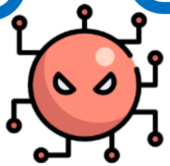
Robustness is model performance a

A life-threatening consequence in safety-critical applications

AI perception is vulnerable to physical-world adversarial attacks

Deep learning (DL) has revolutionized environment perception for artificial autonomous systems such as cars or robots. It is used, for instance, for detecting objects such as pedestrians, vehicles, traffic lights, and traffic signs or for segmenting the drivable area in a scene. Deep neural networks are increasingly embedded into sensors such as the Bosch Multi-Purpose Camera, enabling future applications involving video-based driver assistance systems or highly automated driving.

Environment perception is crucial for autonomous systems.



Attacker

FGSM[1] $\hat{\boldsymbol{x}} = \boldsymbol{x} + \epsilon \times sgn(\nabla_x J(g(x, \theta), t))$

PGD[2] $\hat{x}^k$

IEEE Spectrum | How Adversarial Attacks Could Destabilize Military AI Systems | Type to search

NEWS | ARTIFICIAL INTELLIGENCE

How Adversarial Attacks Could Destabilize Military AI Systems › Adversarial attacks threaten the safety of AI and robotic technologies. Can we stop them?

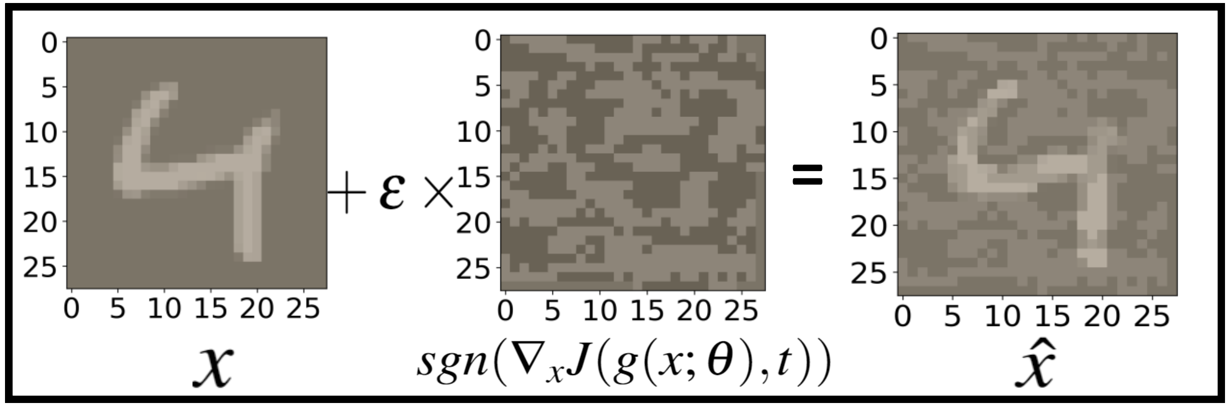
BY DAVID DANKS | 26 FEB 2020 | 6 MIN READ

[1] *Ian J. Goodfellow et al.,"Explaining and harnessing adversarial examples", ICLR 2014.*
[2] *Aleksander Madry et al., "Towards Deep Learning Models Resistant to Adversarial Attacks, ICLR 2018*

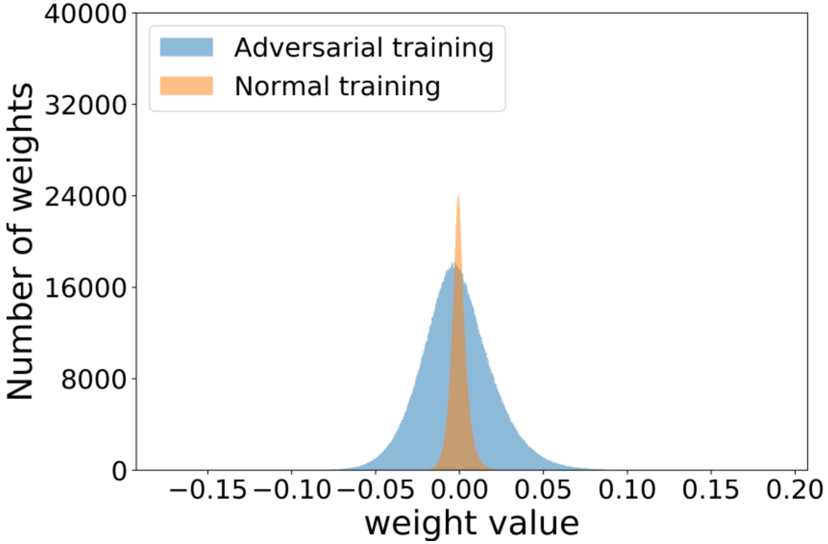
Adversarial Training Demands More Weights

Clean image

Perturbed image



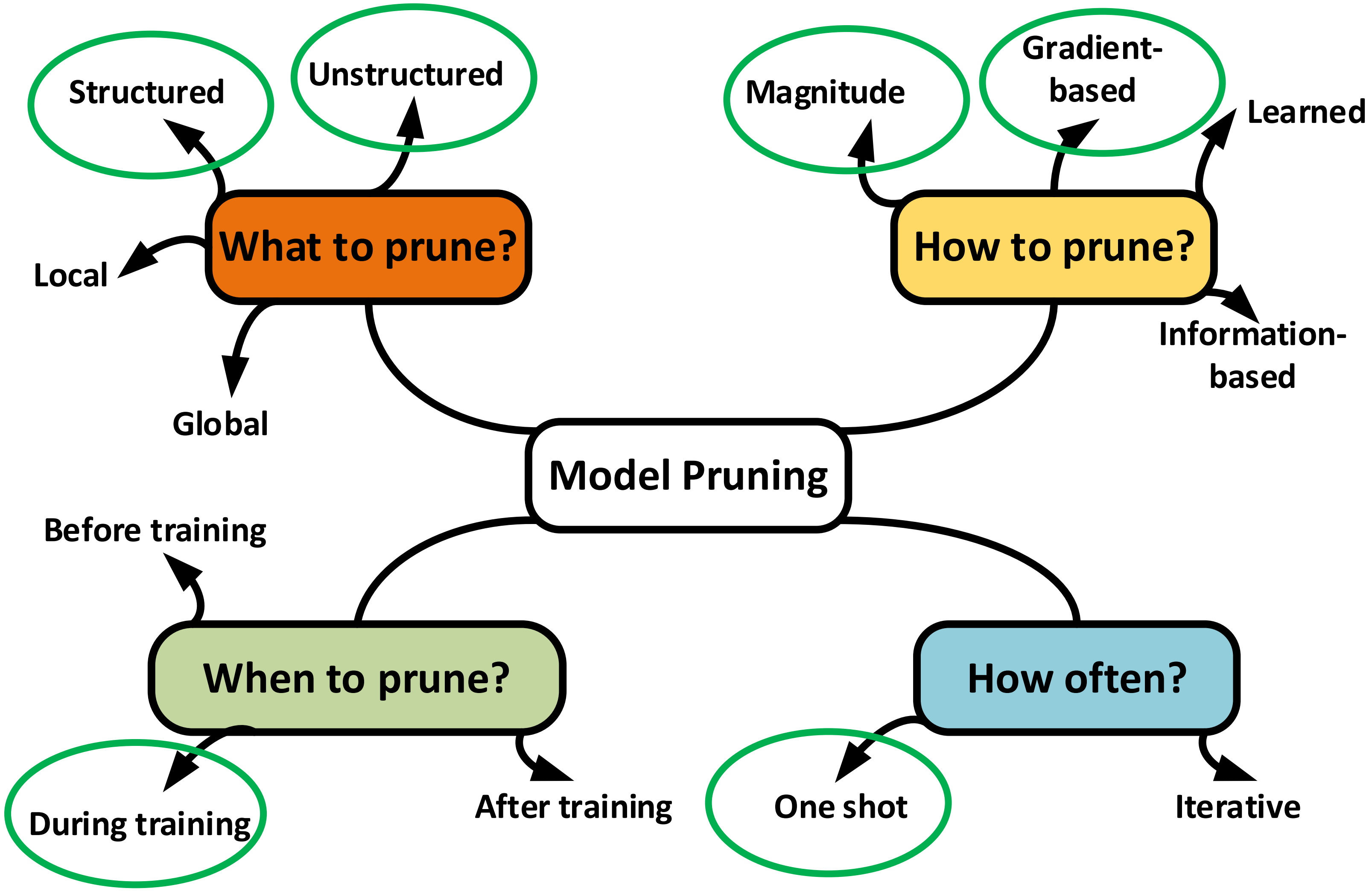
Train with both



Hence, pruning is difficult

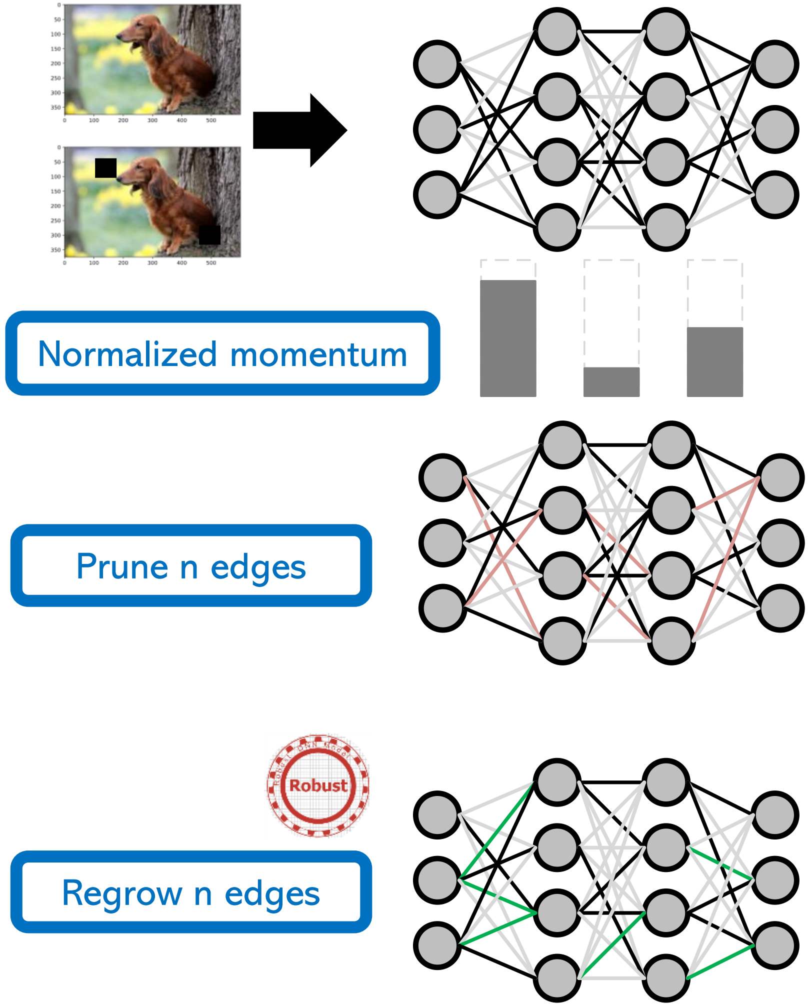
Souvik Kundu et al., "DNR: A Tunable Robust Pruning Framework through Dynamic Network Rewiring of DNNs, ASP-DAC 2021.

Our Unified Pruning Solution: Overview



Souvik Kundu et al., "DNR: A Tunable Robust Pruning Framework through Dynamic Network Rewiring of DNNs, ASP-DAC 2021.

Robust Dynamic Network Rewiring (DNR)



Calculate momentum distribution per layer

Prune fraction of smallest weights from each layer

Redistribute edges according to weights having larger momentums

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

Newly removed edges

Newly regrown edges

Souvik Kundu et al., "DNR: A Tunable Robust Pruning Framework through Dynamic Network Rewiring of DNNs, ASP-DAC 2021.

DNR: Loss Components

USC University of Southern California

$$J_{tot} = \underbrace{\beta\mathcal{L}_{\rho}(\boldsymbol{\theta}, \mathbf{z}, \mathbf{m})} + (1-\beta)J(g(\hat{\boldsymbol{x}}; \boldsymbol{\theta}, \mathbf{m}), \boldsymbol{t})$$

$$\mathcal{L}_{\rho}(\boldsymbol{\theta}, \mathbf{z}, \mathbf{m}) = J(g(\boldsymbol{x}; \boldsymbol{\theta}, \mathbf{m}), \boldsymbol{t}) + \frac{\rho}{2}\sum_{l=1}^{L}||\boldsymbol{\theta}_l \odot \mathbf{m}_l - \mathbf{z}_l||_2^2$$

Model	Method: DNR	Accuracy (%) with irregular pruning			Accuracy (%) with channel pruning		
		Clean	FGSM	PGD	Clean	FGSM	PGD
VGG16	Without dynamic L_2	87.01	50.09	40.62	86.28	49.49	41.25
	With dynamic L_2	86.74	52.92	43.21	85.83	51.03	42.36
ResNet18	Without dynamic L_2	87.45	53.52	45.33	87.97	53.10	45.91
	With dynamic L_2	87.32	55.13	47.35	87.49	56.09	48.33

How important is this term?

Souvik Kundu et al., "DNR: A Tunable Robust Pruning Framework through Dynamic Network Rewiring of DNNs, ASP-DAC 2021.

DNR: Support for Channel Pruning

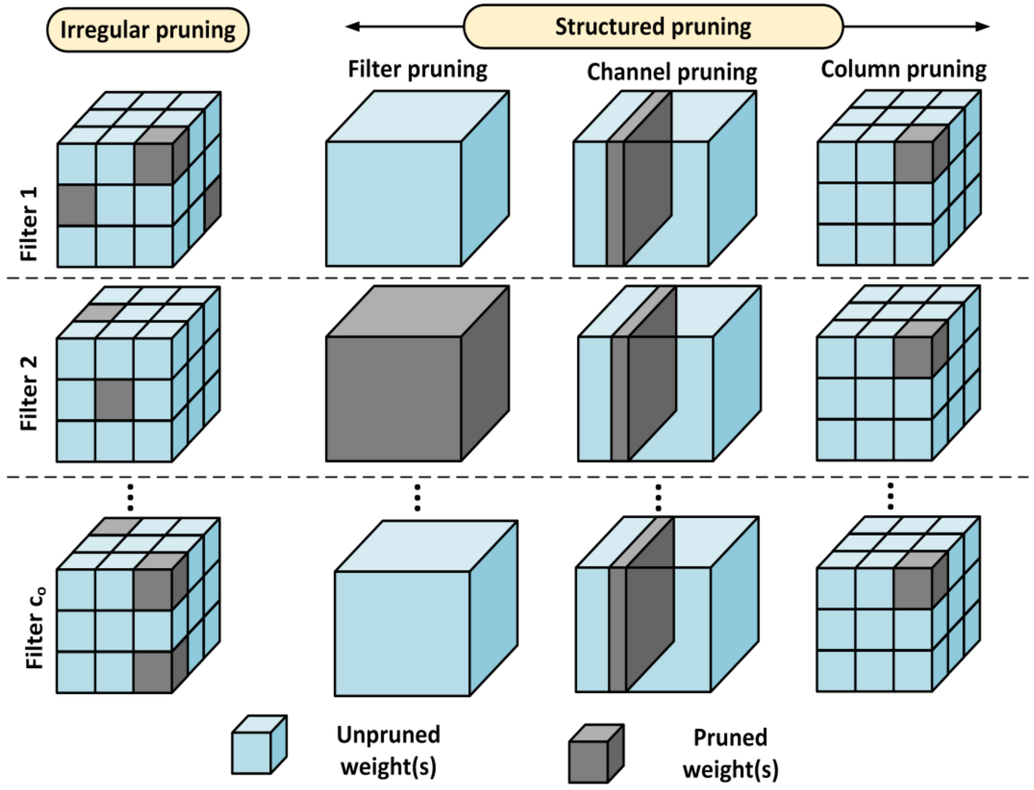
We rank the layer channels based on the L_2 norm of the channel weights

We perform both pruning and regrowing in terms of the granularity of channels instead of weight scalars

Why structured pruning is beneficial for speed-up?

```
for(row=0; row<R; row++) {
 for(col=0; col<C; col++) {
  for(to=0; to<M; to++) {
   for(ti=0; ti<N; ti++) {
    for(i=0; i<K; i++) {
     for(j=0; j<K; j++) {
   L:   output_fm[to][row][col] +=
          weights[to][ti][i][j]*
          input_fm[ti][S*row+i][S*col+j];
} } } } } }
```

N' < N



Souvik Kundu et al., "DNR: A Tunable Robust Pruning Framework through Dynamic Network Rewiring of DNNs, ASP-DAC 2021.

DNR: Why this Approach is Better than SOTA

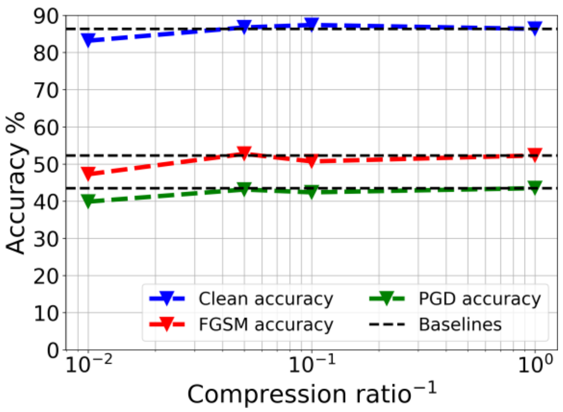
SOTA	DNR	Impact
Iterative	One-shot	Reduced training time
Need per-layer pruning	Decides on the fly	Reduced hyperparameter tuning
Generally pruning and robustness considered as separate problem	Joint optimization with a single loss formulation	Achieves better compression while retaining robustness

Souvik Kundu et al., "DNR: A Tunable Robust Pruning Framework through Dynamic Network Rewiring of DNNs, ASP-DAC 2021.

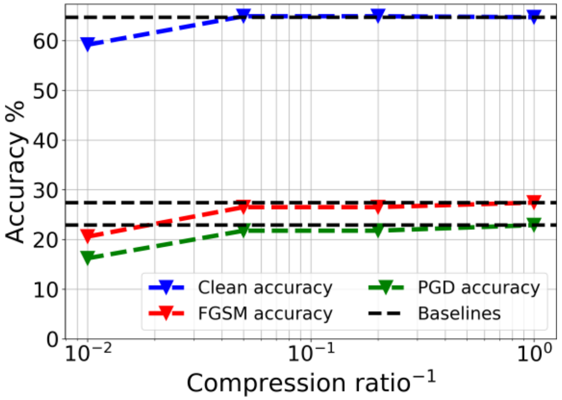
DNR: Compression vs Accuracy trade-off

USC University of Southern California

Irregular Pruning

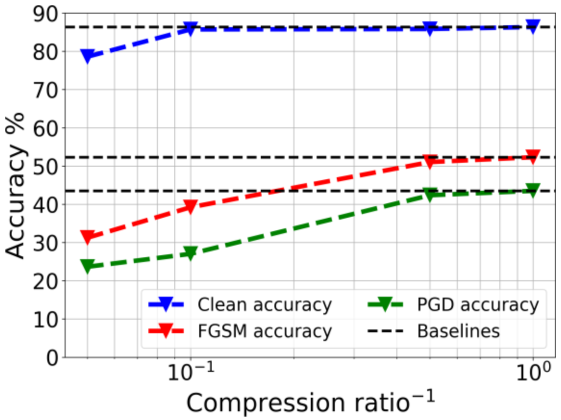


VGG16 on CIFAR-10

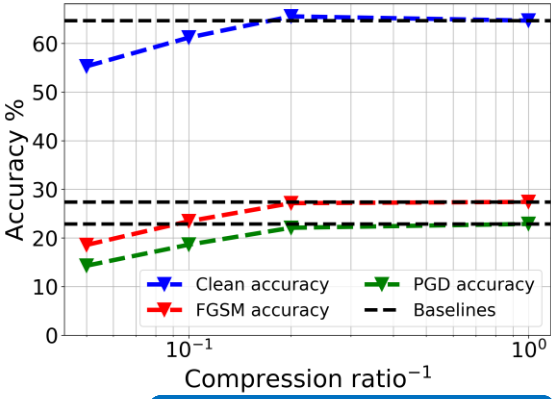


ResNet18 on CIFAR-100

Channel Pruning



VGG16 on CIFAR-10



ResNet18 on CIFAR-100

Channel pruning generally achieves poorer compression than irregular pruning

Souvik Kundu et al., "DNR: A Tunable Robust Pruning Framework through Dynamic Network Rewiring of DNNs, ASP-DAC 2021.

DNR: Comparison with the SOTA

USC University of Southern California

Model	Method	No pre-trained model	Per-layer sparsity knowledge not-needed	Target pruning met	Pruning type	Compre-ssion ratio	Accuracy (%)		
							Clean	FGSM	PGD
VGG16	ADMM [1]	×	×	✓	Irregular	16.78×	86.34	49.52	40.62
	ADMM naive	×	✓	✓		19.74×	83.87	42.46	32.87
	L_1 Lasso [2]	✓	✓	×		*2.01×*	*83.24*	*50.32*	*42.01*
	DNR	✓	✓	✓		**20.85×**	**86.74**	**52.92**	**43.21**
ResNet18	ADMM [1]	×	×	✓	Irregular	14.6×	87.15	54.65	46.57
	ADMM naive	×	✓	✓		19.74×	86.10	50.49	42.24
	L_1 Lasso [2]	✓	✓	×		*6.84×*	*85.92*	***55.20***	*46.80*
	DNR	✓	✓	✓		**21.57×**	**87.32**	55.13	**47.35**

DNR outperforms current SOTA for both clean and perturbed image classification yet maintain increased compression ratio

[1] Ye et al., "Adversarial Robustness vs. Model Compression, or Both?", ICCV 2019.

[2] Rakin et al., "Robust Sparse Regularization: Simultaneously Optimizing Neural Network Robustness and Compactness", GLSVLSI 2020.

Souvik Kundu et al., "DNR: A Tunable Robust Pruning Framework through Dynamic Network Rewiring of DNNs, ASP-DAC 2021.

Summary

- DNR shows a joint adversarial training and sparse learning can yield better compression-robustness trade-off.
- Both structured and irregular pruning can be implemented in the joint training framework of DNR to yield SOTA performance
- Adversarial robustness degrades more rapidly compared to clean image performance for aggressive compression.

1b: Robustness for Brain-inspired Spiking Neural Networks (SNNs)

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

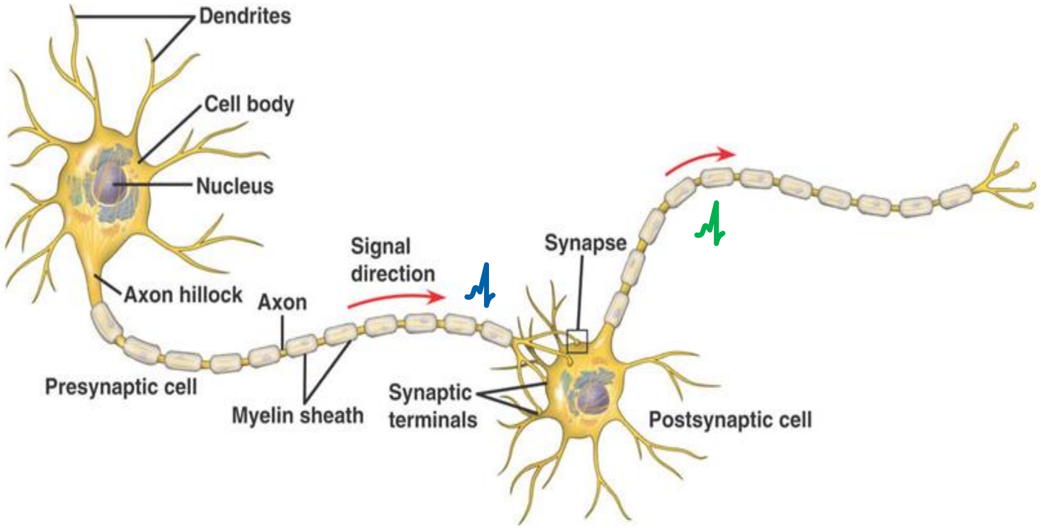
Data and Power Efficient Intelligence with Neuromorphic Learning Machines

Emre O. Neftci[1,2,*]

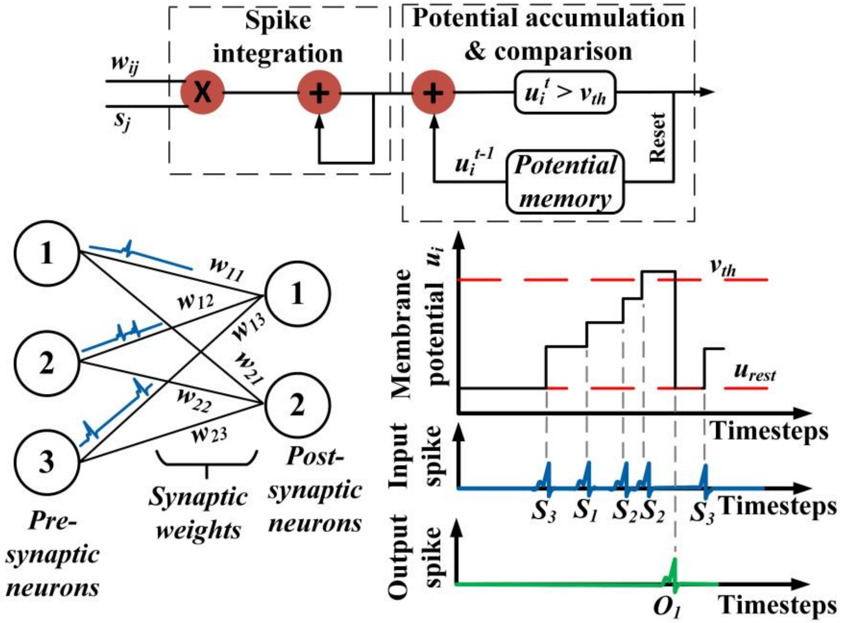
The success of deep networks and recent industry involvement in brain-inspired computing is igniting a widespread interest in neuromorphic hardware that emulates the biological processes of the brain on an electronic substrate. This review explores interdisciplinary approaches anchored in machine learning 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

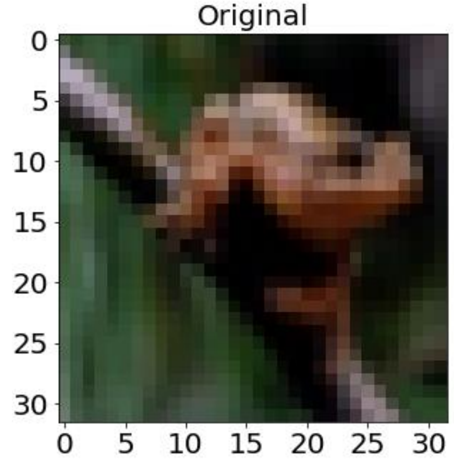
Leaky integrate and fire (LIF) neuron dynamics in discrete time

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

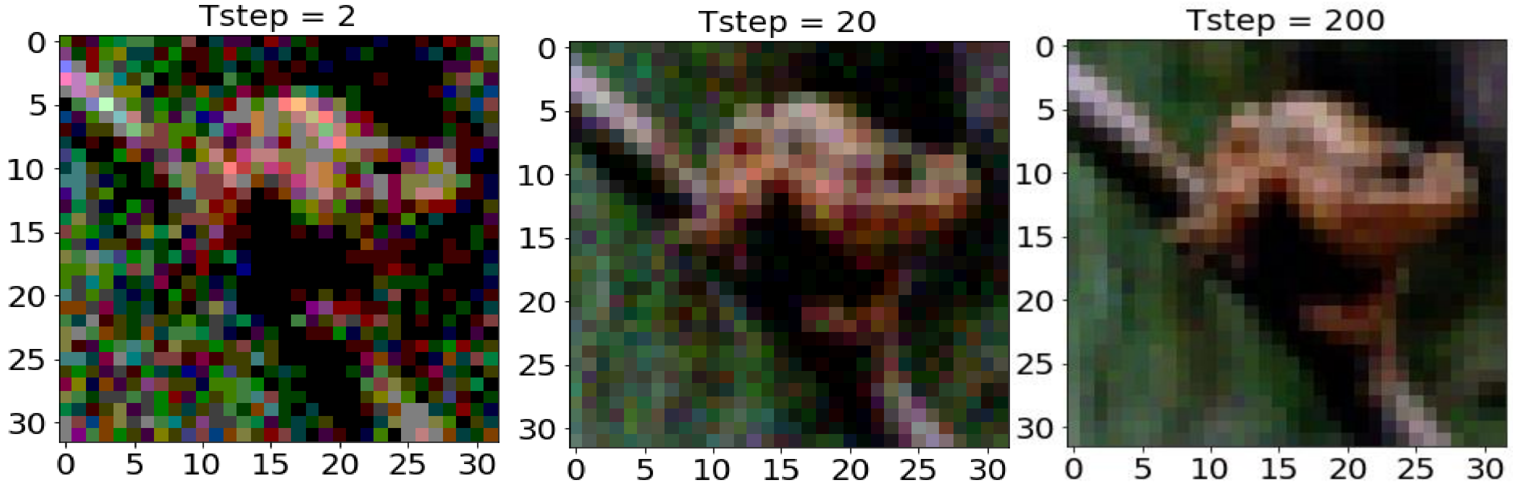
$$O_i^t = \begin{cases} 1, & \text{if } u_i^t > v_{th} \\ 0, & \text{otherwise} \end{cases}$$

- *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.*

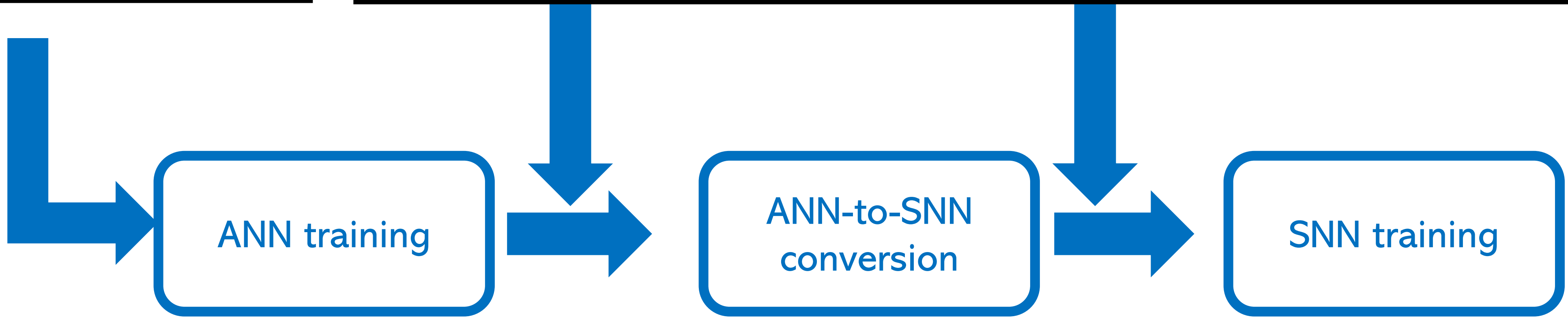
SNN Training Strategy



Original image



Rate-coded input for various timesteps



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

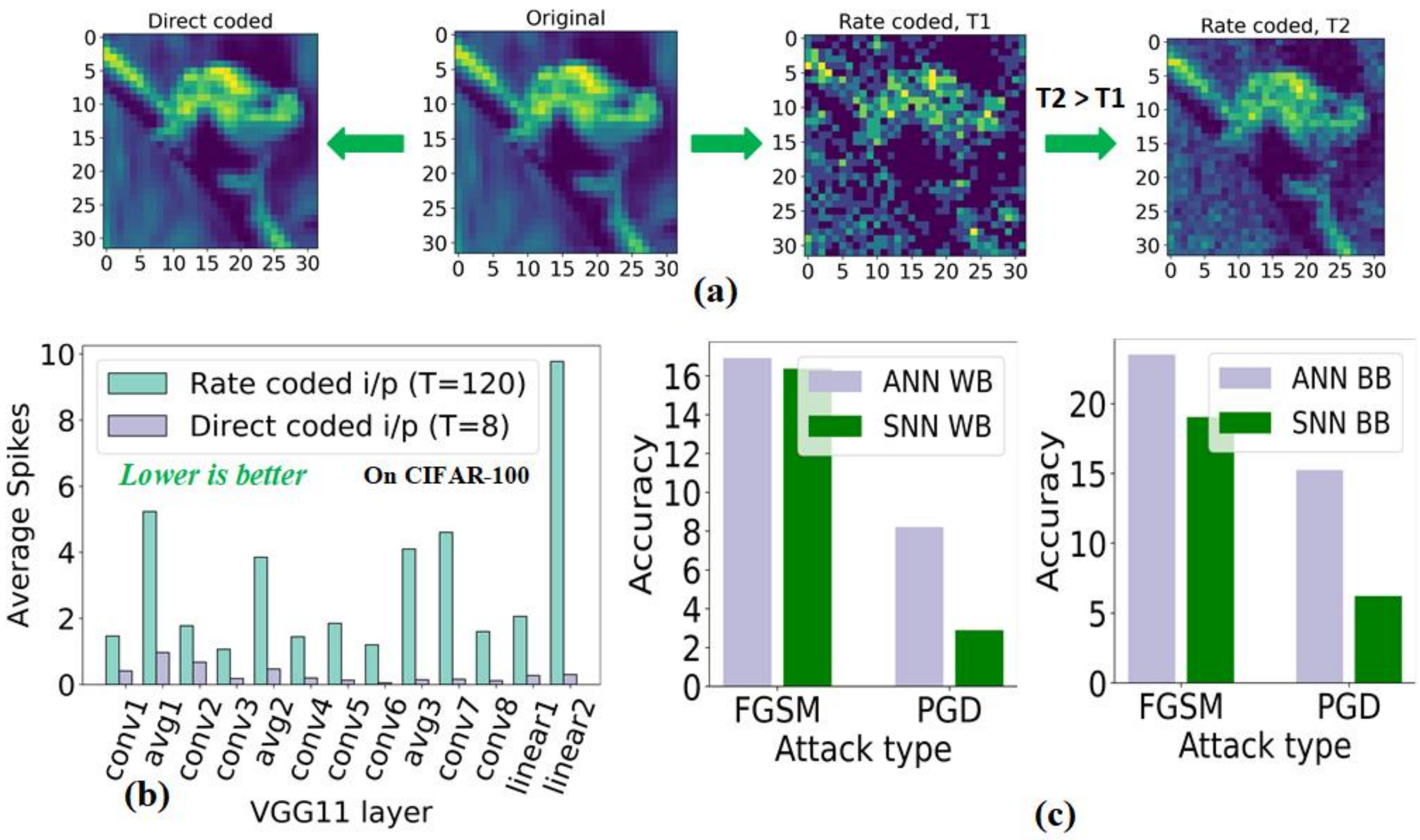
DATE 2021.

- 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

USC University of Southern California

- 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.



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

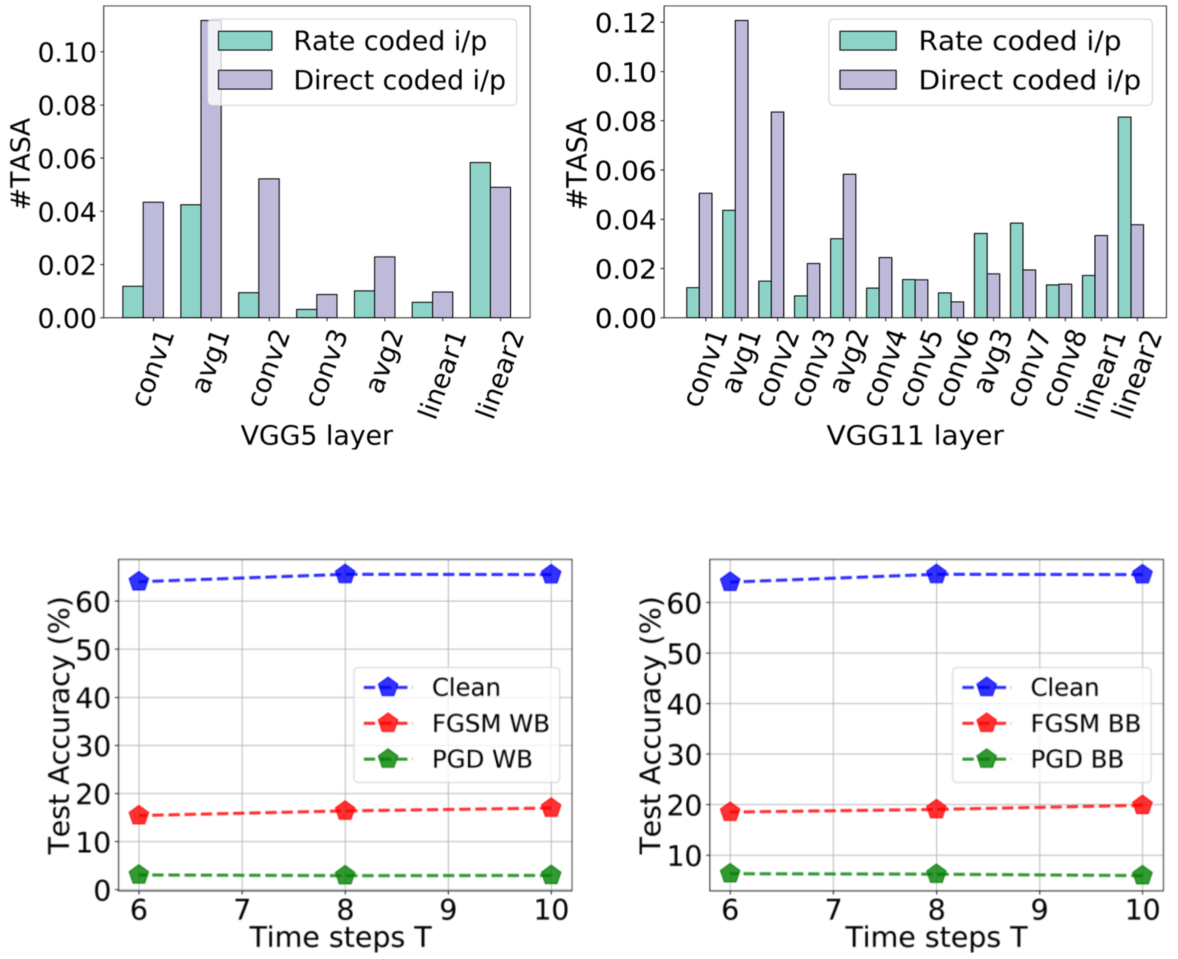
Where do LLSNNs Differ from the Rate-coded Ones?

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

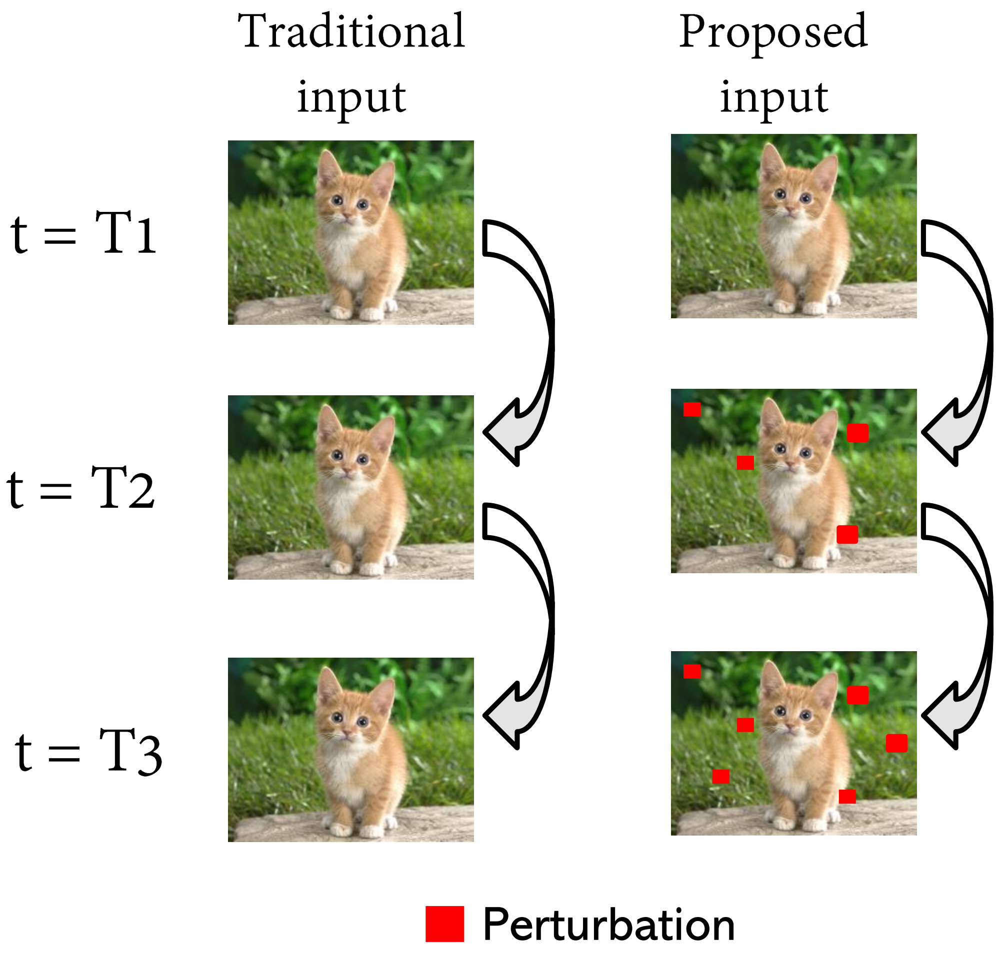
- Reduction in time-step helps improve robustness. However, LLSNNs **can't gain** from further reduction in t-steps.



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

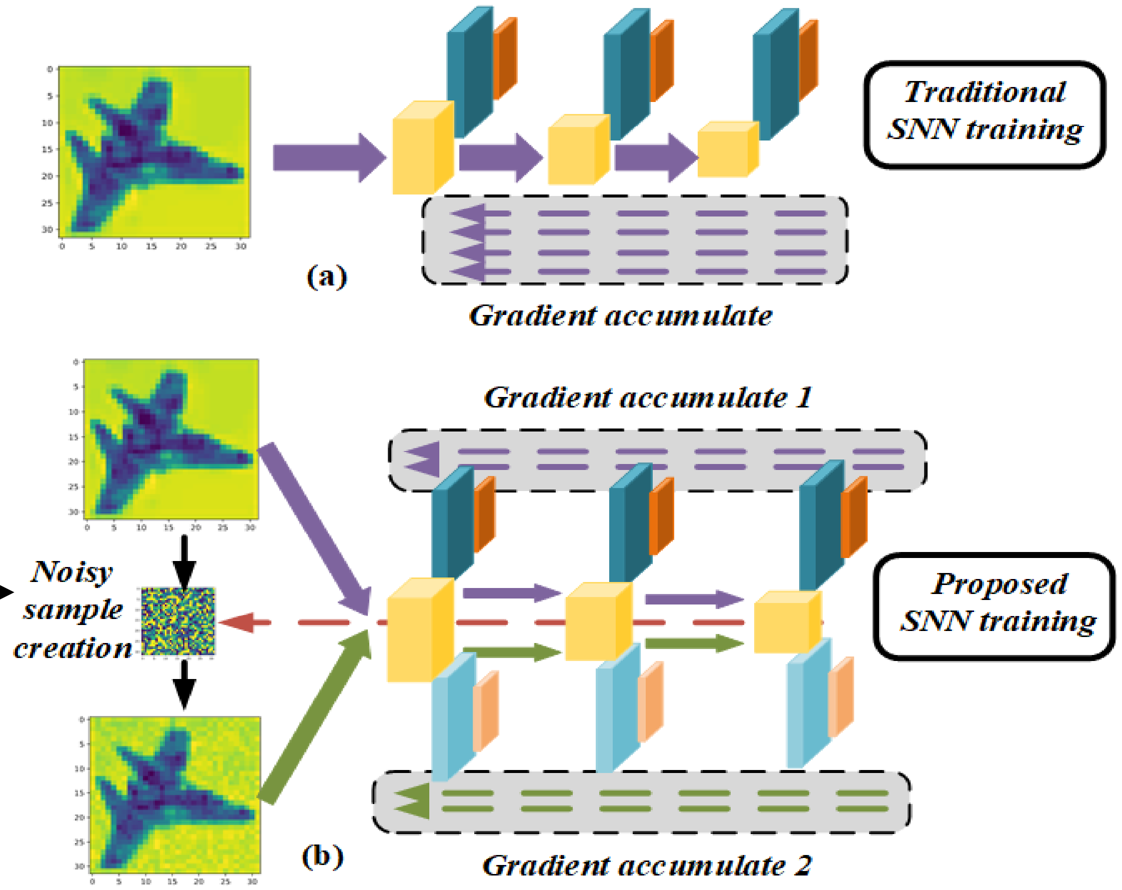
Achieving Robustness for SNNs: HIRE-SNN

- 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.



Souvik 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



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

Souvik 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

USC University of Southern California

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

Souvik 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.

2: Model Privacy Under Distillation

Machine Learning as a Service (MLAAS) is on the Rise

Household robots

Autonomous driving

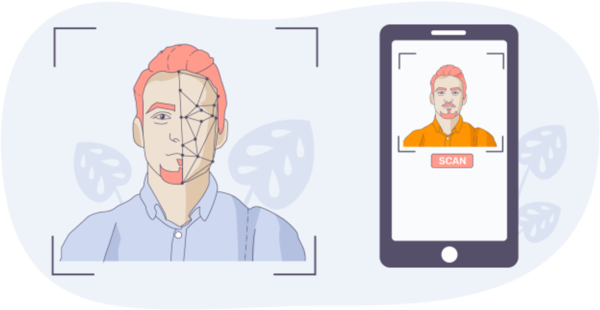


Image analysis

Image courtesy: Google images

- Various trained models are deployed at the edge to perform complex computer vision and natural language processing tasks
- Industries prefer the trained models to be released as commercial black-box APIs

Model Performance Protection is Important

USC University of Southern California

- ➢ Winning teams of AI competitions do **<u>not</u>** want their model performance to be replicated by opponents

- ➢ Industry releasing models as commercial black-box API do **<u>not</u>** want their model performance to be replicated by a potential competitor

- ➢ Commercial black-box ML APIs often require **<u>large</u>** human resource and training costs that the owner wants to be compensated for via MLAAS earnings

Neural Networks

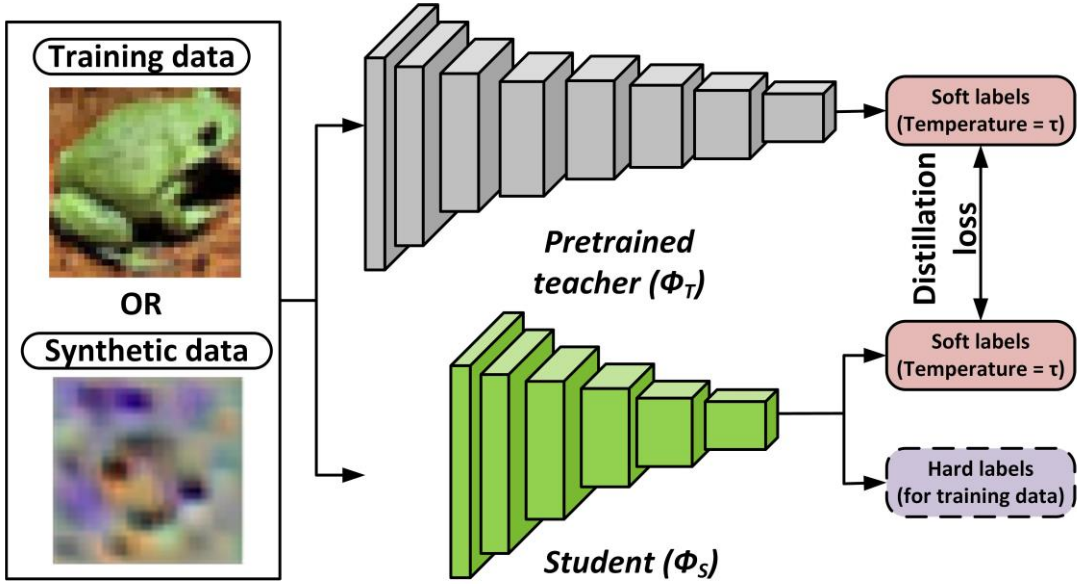
Apply cutting-edge research to train deep neural networks on problems ranging from perception to control. Our per-camera networks analyze raw images to perform semantic segmentation, object detection and monocular depth estimation. Our birds-eye-view networks take video from all cameras to output the road layout, static infrastructure and 3D objects directly in the top-down view. Our networks learn from the most complicated and diverse scenarios in the world, iteratively sourced from our fleet of nearly 1M vehicles in real time. A full build of Autopilot neural networks involves 48 networks that take 70,000 GPU hours to train 🔥. Together, they output 1,000 distinct tensors (predictions) at each timestep.

Source: https://www.tesla.com/AI

Knowledge-Distillation (KD): A Potential Threat to MLAAS

USC University of Southern California

Primary application: model compression



Concerning application: mimicking performance from black-box models

- KD can transfer the “rich” knowledge of a compute-heavy teacher to a compute-efficient student model under both data-available[1] and data-free scenarios[2]

[1] Geoffrey Hinton et al., “Distilling the knowledge in a neural network”, NeurIPS 2014 (workshop).
[2] Paul Micaelli and Amos Storkey, “Zero-shot knowledge transfer via adversarial belief matching”, NeurIPS 2019.

Undistillable Models[1]

USC University of Southern California

- A class of models that
  - Perform similar to standard teacher models to maintain their own performance
  - However, act as “**nasty**” teachers to any student model by not allowing it to mimic performance.
- Core idea
  - Inject **false** sense of generalization to the student[1]

Training loss of Undistillable models (***Φ**$_T$*):

$$\boxed{\mathcal{L}_N = \underbrace{\mathcal{L}_{\mathcal{CE}}\big(\sigma(g_{\Phi_T}(\boldsymbol{x},\boldsymbol{y}))\big)}_{\text{Cross-entropy (CE) loss}} - \underbrace{\alpha_N * \tau_N^2 * \mathcal{L}_{\mathcal{KL}}\big(\sigma(g_{\Phi_T}(\boldsymbol{x},\boldsymbol{y}),\tau_N),\sigma(g_{\Phi_A}(\boldsymbol{x},\boldsymbol{y}),\tau_N)\big)}_{\text{Self-undermining loss}}}$$

[1] Haoyu Ma et al., “Undistillable: Making a nasty teacher that cannot teach students”, ICLR 2021 (spotlight).

A1: Analyzing Undistillability

USC University of Southern California

➢ A study of transferability of the impact of nasty teachers

Teacher	Teacher type	Teacher Acc %	Student Acc %	Δ_{base}
ResNet50	Nasty	76.57	72.47	-5.08
ResNet18	Distilled	$\boxed{72.47}$	$\boxed{70.99}$	-6.56
ResNet50	Normal	78.04	79.39	+1.84
ResNet18	Distilled	79.39	79.47	+1.92

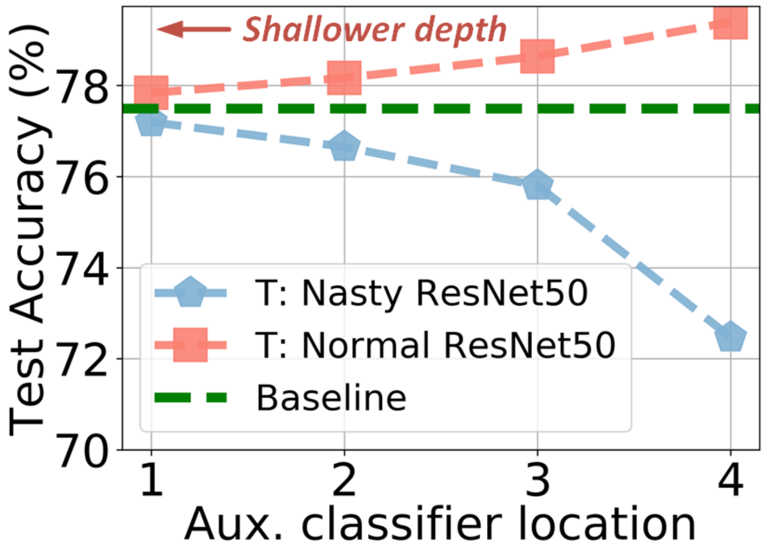
The nastiness of a teacher transfers to its student

Souvik Kundu et al., "Analyzing the Confidentiality of Undistillable Teachers in Knowledge Distillation", NeurIPS 2021.

A2: Analyzing the Undistillability

USC University of Southern California

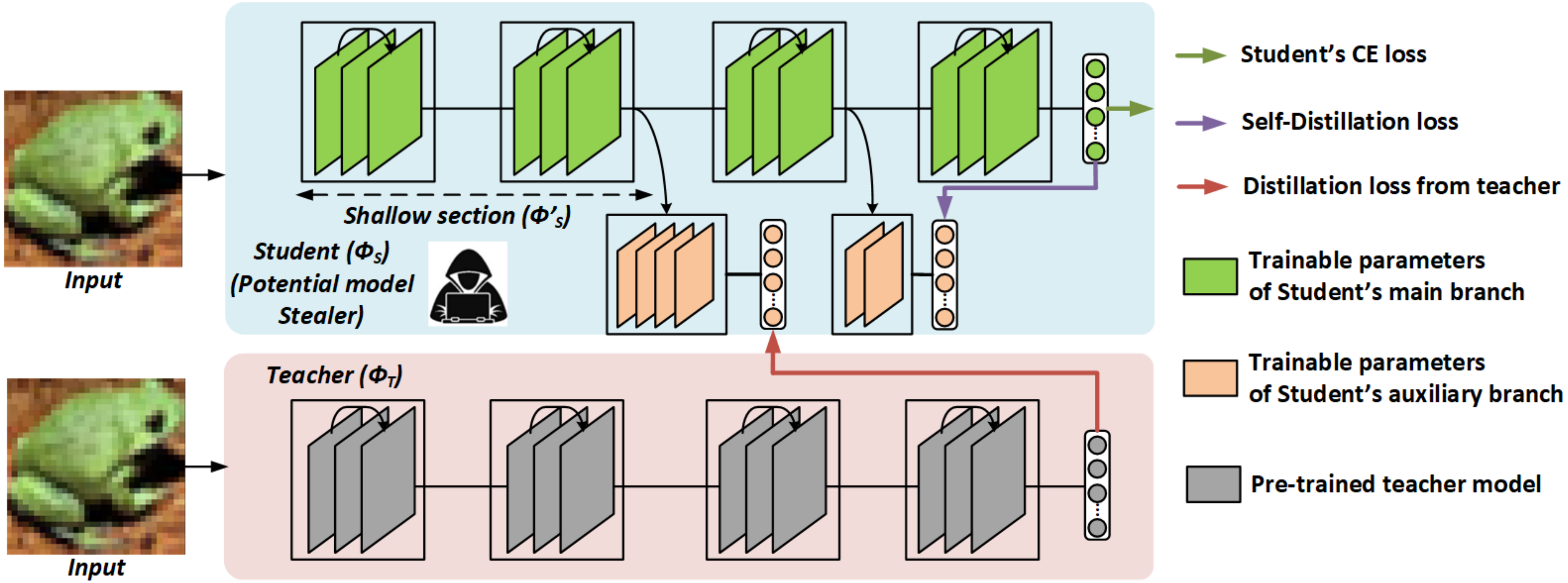
➢ A study of applying KD at various depth of the student model



Impact of a teacher reduces as we use KD at shallower depths of student

Souvik Kundu et al., "Analyzing the Confidentiality of Undistillable Teachers in Knowledge Distillation", NeurIPS 2021.

Our Proposal: Skeptical Student



- Transfer knowledge to shallow depth ($\boldsymbol{\Phi'_S}$) of a student via aux. classifier (AC)
- Use self-distillation at AC in $\boldsymbol{\Phi_S}$- $\boldsymbol{\Phi'_S}$ to boost performance of student $\boldsymbol{\Phi_S}$

Souvik Kundu et al., "Analyzing the Confidentiality of Undistillable Teachers in Knowledge Distillation", NeurIPS 2021.

Skeptical Students: Training Loss

KL-divergence loss component:

$$\mathcal{L}_T = (1-\alpha) * \mathcal{L}_{\mathcal{CE}}\big(\sigma(g_{\Phi'_S}(\boldsymbol{x},\boldsymbol{y}))\big) + \alpha * \tau^2 * \mathcal{L}_{\mathcal{KL}}\big(\sigma(g_{\Phi'_S}(\boldsymbol{x},\boldsymbol{y}),\tau),\sigma(g_{\Phi_T}(\boldsymbol{x},\boldsymbol{y}),\tau)\big)$$

Self-distillation loss component :

$$\mathcal{L}_{SD} = \sum_{j\in\mathcal{J}} \{(1-\beta) * \mathcal{L}_{\mathcal{CE}}\big(\sigma(g_{\Phi^j_S}(\boldsymbol{x},\boldsymbol{y}))\big) + \beta * \mathcal{L}_{\mathcal{KL}}\big(\sigma(g_{\Phi^j_S}(\boldsymbol{x},\boldsymbol{y}),\tau),\sigma(g_{\Phi_S}(\boldsymbol{x},\boldsymbol{y}),\tau)\big)\}$$

CE loss component :

$$\mathcal{L}_{\mathcal{CE}}\big(\sigma(g_{\Phi_S}(\boldsymbol{x},\boldsymbol{y}))\big)$$

Total loss (hybrid distillation):

$$\mathcal{L}_S = \gamma_1\mathcal{L}_T + \gamma_2\mathcal{L}_{SD} + \gamma_3\mathcal{L}_{\mathcal{CE}}\big(\sigma(g_{\Phi_S}(\boldsymbol{x},\boldsymbol{y}))\big)$$

Souvik Kundu et al., "Analyzing the Confidentiality of Undistillable Teachers in Knowledge Distillation", NeurIPS 2021.

Skeptical Students: Distilled from Nasty Teachers

USC University of Southern California

Dataset	Φ_T	Φ_T Acc. (%)	Φ_S	Φ_S Baseline Acc. (%)	Student Acc. (%) Normal (acc_n)	Skeptical (acc_s)	Skeptical-E (acc_{se})	Δ_{acc}
CIFAR-10	ResNet18	94.67	ResNet18	95.15	94.13(±0.18)	**95.09**(± 0.15)	94.77(± 0.05)	+0.96
			MobileNetV2	90.12	88.13(±0.13)	**90.37**(± 0.25)	90.21(± 0.18)	+2.24
	ResNet50	94.28	ResNet18	95.15	94.38(±0.18)	**95.16**(± 0.01)	95.02(± 0.01)	+0.78
			ResNet50	94.9	94.21(±0.04)	**95.48**(± 0.14)	95.48(± 0.14)	+1.27
			MobileNetV2	90.12	88.76(±0.14)	**91.02**(± 0.09)	90.88(± 0.23)	+2.26
CIFAR-100	ResNet18	77.55	ResNet18	77.55	75.00(±0.14)	**77.33**(± 0.21)	76.38(± 0.1)	+2.33
			MobileNetV2	69.24	7.13(±0.71)	**66.62**(± 0.30)	64.26(± 0.64)	+59.49
	ResNet50	76.57	ResNet18	77.55	72.28(±0.27)	**77.25**(± 0.25)	75.48(±0.54)	+4.97
			ResNet50	78.04	74.14(±0.85)	**78.65**(± 0.29)	77.61(±0.1)	+4.52
			MobileNetV2	69.24	7.72(±1.57)	**66.38**(± 0.50)	62.93(±0.75)	+58.66
Tiny-ImageNet	ResNet18	62.08	ResNet18	63.07	53.60(±0.04)	**65.76**(±0.83)	60.63(±0.07)	+12.16
			MobileNetV2	57.01	4.81(± 0.19)	**54.74**(±0.84)	54.27(±2.94)	+49.93

Skeptical students achieve similar to teacher performance **even when the teacher is Undistillable** (or nasty).

Souvik Kundu et al., "Analyzing the Confidentiality of Undistillable Teachers in Knowledge Distillation", NeurIPS 2021.

Skeptical Students: Distilled from Normal Teachers

USC University of Southern California

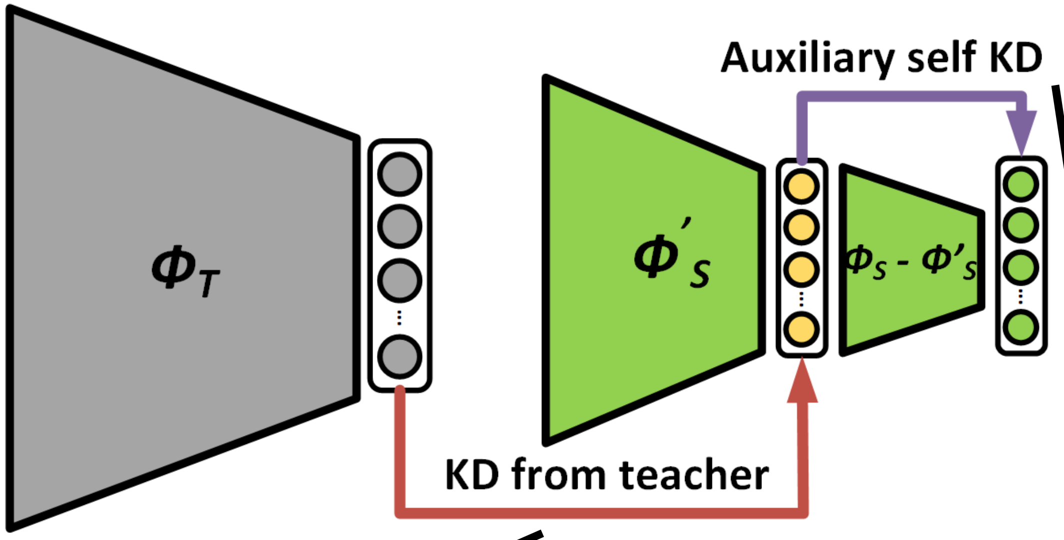
Dataset	Φ_T	Φ_T Acc. (%)	Φ_S	Φ_S Base-line Acc. (%)	Student Acc. (%) Normal (acc_n)	Skeptical (acc_s)	Skeptical-E (acc_{s_e})	Δ_{acc}
CIFAR-10	ResNet18	95.15	ResNet18	95.15	95.38 ($\pm$0.10)	**95.45**($\pm$0.10)	95.42($\pm$0.09)	+0.07
			MobileNetV2	90.12	91.36($\pm$0.17)	91.81($\pm$0.15)	**92.00**($\pm$0.28)	+0.64
	ResNet50	94.9	ResNet18	95.15	**95.43**($\pm$0.11)	95.31($\pm$0.01)	95.27($\pm$0.04)	-0.12
			ResNet50	94.9	95.15($\pm$0.13)	95.85($\pm$0.05)	**96.09**($\pm$0.01)	+0.94
			MobileNetV2	90.12	91.71($\pm$0.06)	91.71($\pm$0.18)	**91.95**($\pm$0.16)	+0.24
CIFAR-100	ResNet18	77.55	ResNet18	77.55	78.96($\pm$0.12)	78.79($\pm$0.42)	**79.68**($\pm$0.52)	+0.72
			MobileNetV2	69.24	75.12($\pm$0.08)	71.63($\pm$0.19)	**75.45**($\pm$0.06)	+0.33
	ResNet50	78.04	ResNet18	77.55	79.21($\pm$0.24)	78.51($\pm$0.44)	**79.86**($\pm$0.01)	+0.65
			ResNet50	78.04	79.56($\pm$0.13)	80.66($\pm$0.52)	**81.96**($\pm$0.52)	+2.4
			MobileNetV2	69.24	75.28($\pm$0.04)	71.76($\pm$0.16)	**76.32**($\pm$0.34)	+1.04
Tiny-ImageNet	ResNet18	63.07	ResNet18	63.07	67.35($\pm$0.18)	66.49($\pm$0.30)	**67.43**($\pm$0.47)	+0.08
			MobileNetV2	57.01	64.99($\pm$0.51)	59.37($\pm$0.01)	**65.38**($\pm$0.01)	+0.39

Skeptical students achieve **similar to normal students' performance** upon distillation from a normal teacher.

Souvik Kundu et al., "Analyzing the Confidentiality of Undistillable Teachers in Knowledge Distillation", NeurIPS 2021.

Skeptical Students: Data-free Distillation

USC University of Southern California



$$\mathcal{L}_{S_{DF}} = \mathcal{L}_{\mathcal{KL}}\big(\sigma(g_{\Phi'_S}(\boldsymbol{x},\boldsymbol{y}),\tau),\sigma(g_{\Phi_T}(\boldsymbol{x},\boldsymbol{y}),\tau)\big) + \mathcal{L}_{\mathcal{KL}}\big(\sigma(g_{\Phi_S}(\boldsymbol{x},\boldsymbol{y}),\tau),\sigma(g_{\Phi'_S}(\boldsymbol{x},\boldsymbol{y}),\tau)\big) + \boxed{\gamma_{at}\mathcal{L}_{\mathcal{AT}}}$$

Grey-box teacher: Attention transfer (AT) loss ✔

Black-box teacher: Attention transfer (AT) loss ✘

Souvik Kundu et al., "Analyzing the Confidentiality of Undistillable Teachers in Knowledge Distillation", NeurIPS 2021.

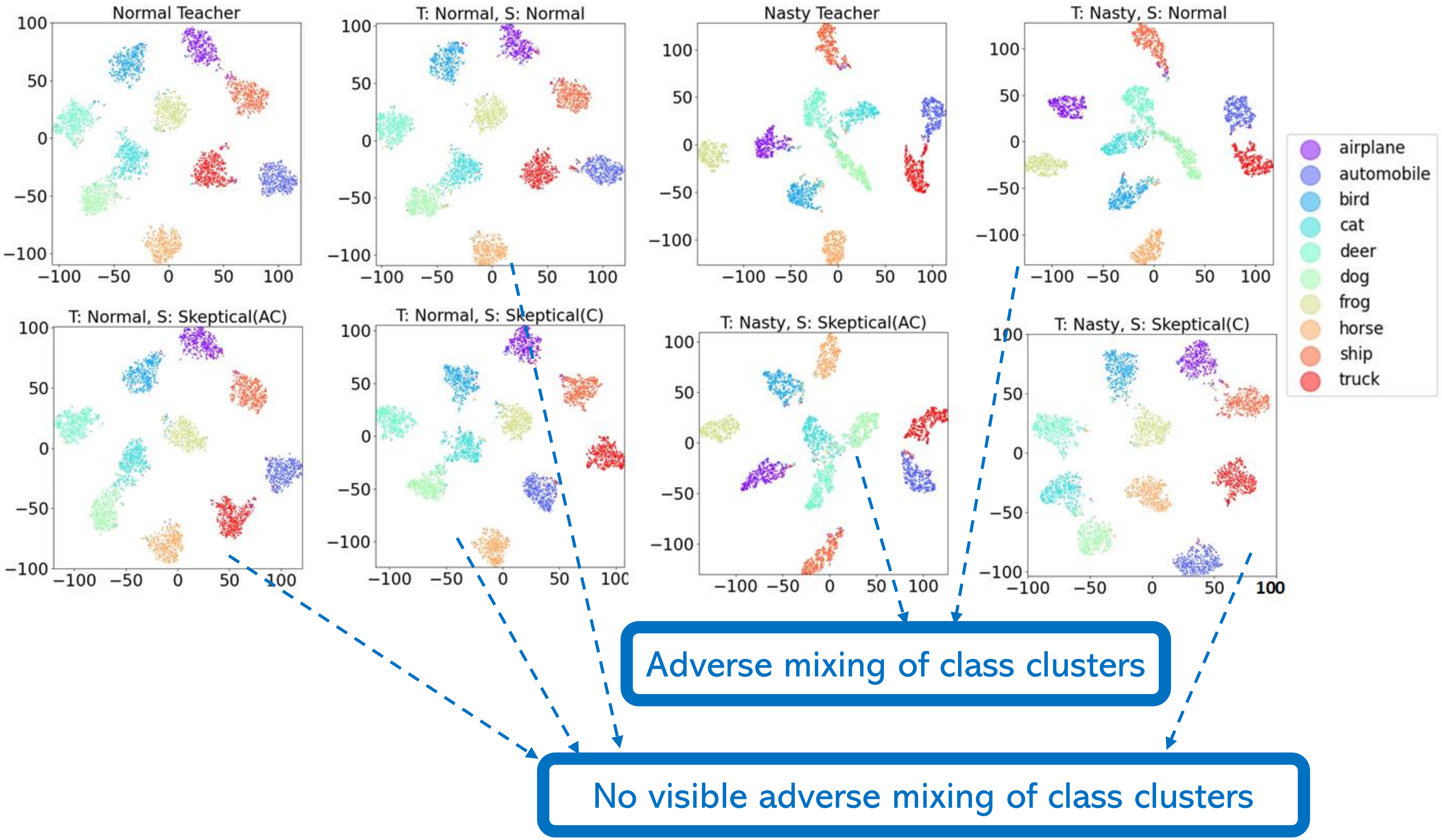
Skeptical Students: Data-free Distillation Results

USC University of Southern California

Dataset	Φ_T	Φ_T type	Φ_T Acc. (%)	Φ_S	Student Acc. (%) Normal	Student Acc. (%) Skeptical	Δ_{acc}
With AT loss (grey-box)							
CIFAR -10	ResNet34	Nasty	94.81	ResNet18	87.7($\pm$1.20)	**91.76**($\pm$0.30)	+4.06
		Normal	95.3		93.41($\pm$0.21)	**93.52**($\pm$0.06)	+0.11
	ResNet50	Nasty	94.28		80.34($\pm$1.19)	**86.14**($\pm$0.01)	+5.80
		Normal	94.9		90.54($\pm$1.16)	**91.93**($\pm$0.04)	+1.39
Without AT loss (black-box)							
CIFAR -10	ResNet50	Nasty	94.28	ResNet18	20.95($\pm$0.21)	**79.93**($\pm$1.58)	**+58.98**
		Normal	94.9		22.08($\pm$0.56)	**80.71**($\pm$1.21)	+58.63

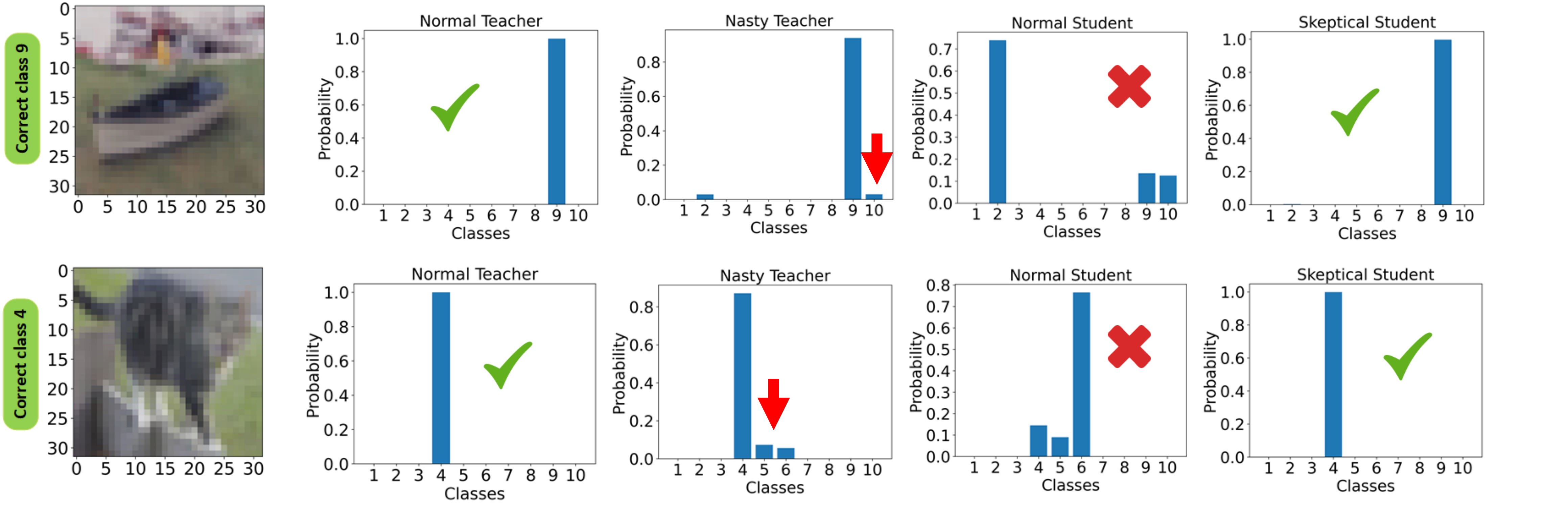
Skeptical students achieve significantly superior performance compared to normal counter parts.

Skeptical Students: Analysis of Results



Evaluations done on CIFAR-10 dataset with ResNet50 as teacher and ResNet18 as student model.

Skeptical Students: Analysis of Results



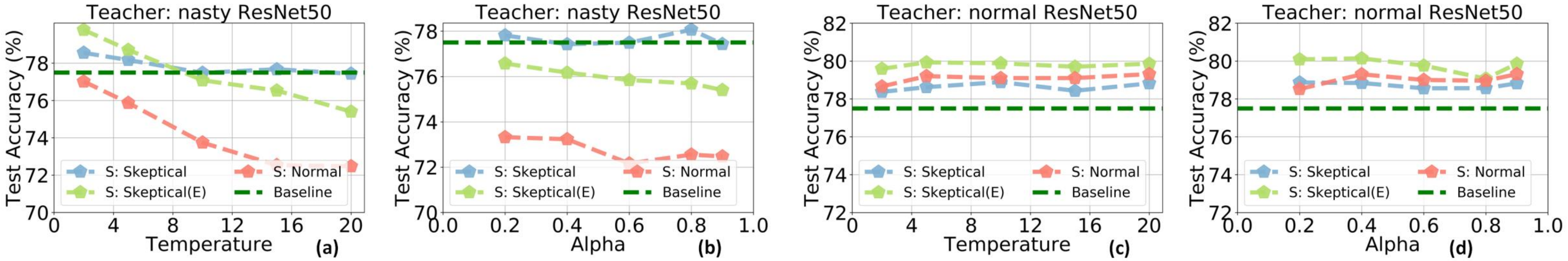
Negligible logit values of incorrect classes

Non-negligible logit values of incorrect classes

Incorrectly classified class

Evaluations done on CIFAR-10 dataset with ResNet50 as teacher and ResNet18 as student model.

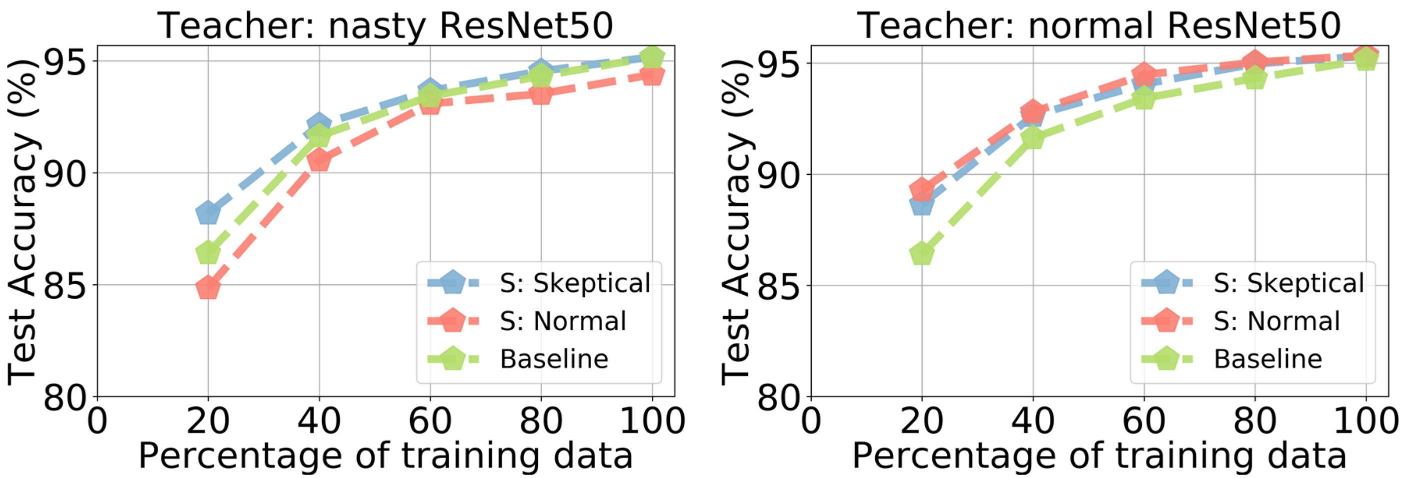
Skeptical Students: Ablation with Hyperparameters



Skeptical students consistently outperform normal counter parts on different loss strength and temperature value choices[1].

[1] *Evaluation done on CIFAR-100 dataset to ResNet18 student model.*

Skeptical Students: Ablation with Limited Data-availability



Skeptical students consistently outperform normal counter parts on various limited data availability scenarios[1].

[1] *Evaluation done on CIFAR-10 dataset to ResNet18 student model.*

Skeptical Students: Transferability of Nastiness

USC University of Southern California

Teacher	Teacher type	Teacher Acc %	Student Acc %	Δ_{base}
ResNet50	Nasty	76.57	77.43	-0.12
ResNet18	Nasty-distilled	77.43	79.22	+1.67
ResNet50	Normal	78.04	78.90	+1.35
ResNet18	Normal-distilled	78.90	79.92	+2.37

The nastiness of a teacher does not get transferred to the skeptical student

Summary

- Skeptical students can successfully **distill from even a nasty teacher** outperforming normal student counterparts

- Skeptical students can yield **better performance** on both data-available and data-free scenarios

- The success of skeptical students in mimicking model performance **poses a fundamental question** on protecting model IP in a distillation framework.

Conclusions

USC University of Southern California

- With the limitation Moore's law and Denard's scaling need for hardware-algorithm co-design has grown a lot.
- With the shift of computing workloads from cloud to edge demand for efficiency, robustness and privacy has grown a lot.
- As an A.I. researcher my goal is to thrive towards an A.I. augmented sustainable, safe and secure future.

Need to understand hardware limitations

Need to understand the societal demand

Need to understand the responsibility

Selected First Author Publications

USC University of Southern California

1. C **[DATE 2022]** ***S. Kundu*** et al., "BMPQ: Bit-Gradient Sensitivity Driven Mixed-Precision Quantization of DNNs from Scratch".
2. C **[NeurIPS 2021]** ***S. Kundu*** et al., "Analyzing the Confidentiality of Undistillable Teachers in Knowledge Distillation".
3. C **[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".
4. C **[CVPRW 2021]** ***S. Kundu*** et al., "Skeptical Student: Diminishing the Effect of Leaking Teacher in Knowledge Distillation".
5. C **[ICASSP 2021]** ***S. Kundu*** et al., "AttentionLite: Towards Efficient Self-Attention Models for Vision".
6. C **[WACV 2021]** ***S. Kundu*** et al., "Spike-Thrift: Towards Energy-Efficient Deep Spiking Neural Networks by Limiting Spiking Activity via Attention-Guided Compression".
7. C **[ASP-DAC 2021]** ***S. Kundu*** et al., "DNR: A Tunable Robust Pruning Framework Through Dynamic Network Rewiring of DNNs".
8. J **[ACM TECS 2022]** ***S. Kundu*** et al., "Towards Adversary aware Non-Iterative Model Pruning Through Dynamic Network Rewiring of DNNs".
9. J **[IEEE TC 2020]** ***S. Kundu*** et al., "Pre-defined Sparsity for Low-Complexity Convolutional Neural Networks".

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

Thank You!

“Success in creating AI would be the biggest event in human history. Unfortunately, it might also be the last, unless we learn how to avoid the risks.”

-- Stephen Hawking