Robust and Efficient Machine Learning for Mission-Critical Applications

Bhavya Kailkhura (SafeML LDRD)

Machine Intelligence Group
Lawrence Livermore National Laboratory

This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under contract DE-AC52-07NA27344. Lawrence Livermore National Security, LLC

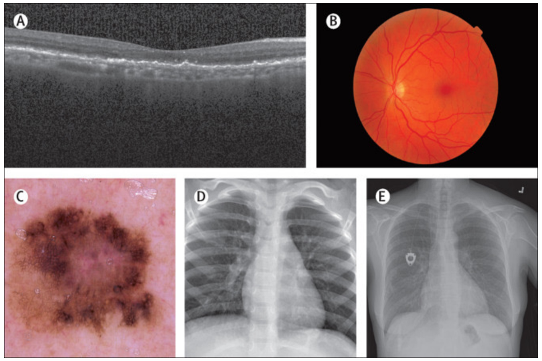
Lawrence Livermore National Laboratory

AI at LLNL/DOE

Cyber-Physical Security

Power Grid

Healthcare



Faulty and slow decisions can risk human safety or incur significant cost
(experimental resources or lost opportunities)

Existing AI techniques are quite brittle

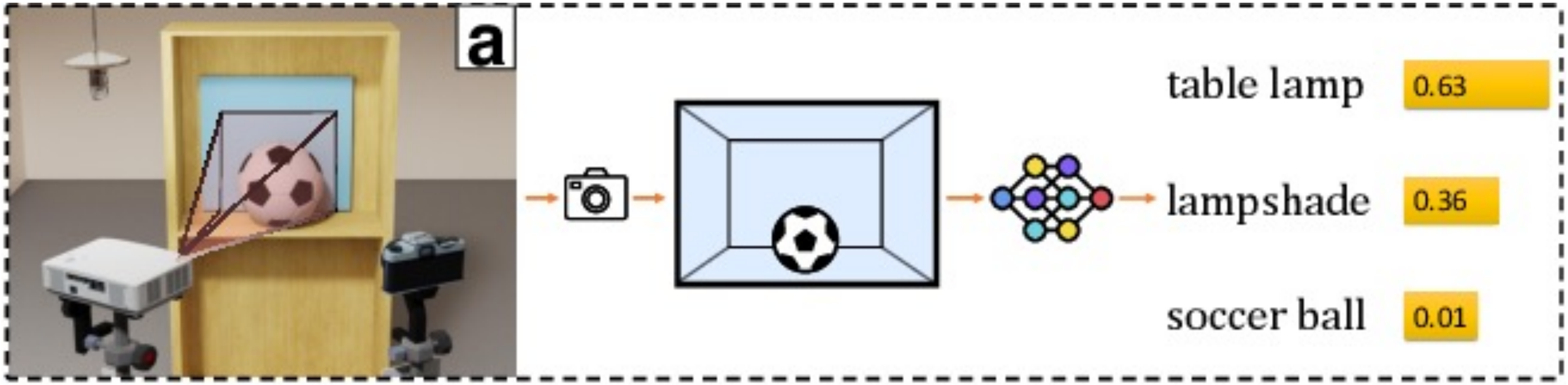
- **Red AI:** Designing an input, which seems normal for a human but is wrongly classified by ML models
  - Applicable to images, text, graphs, etc.
  - Spam filtering, malware detection, intrusion detection, etc.

- Demos:
  - Attacking an image classification system
  - Attacking a text-based search system

LLNL’s Red AI: digital to physical world

- Our **Red AI team** has developed real-time physical world attack to gauge model vulnerabilities

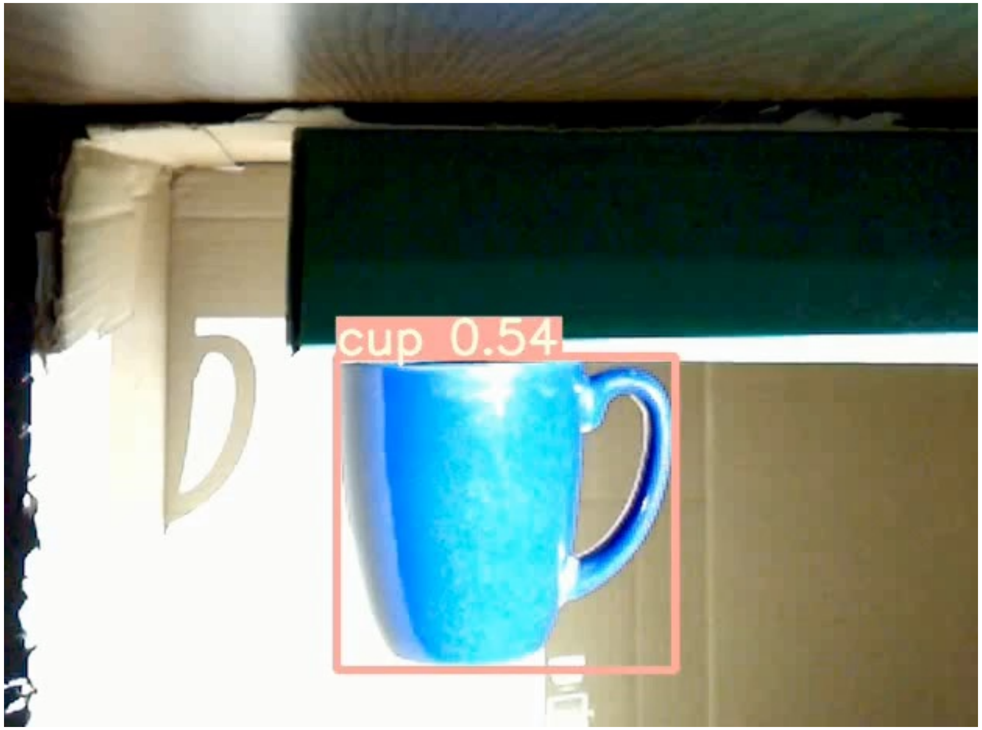
- Our attack algorithm uses a light-projector to fool machine learning based video surveillance systems
  - Applicable to any predictive model, e.g., deep neural nets, random forest, rule-based systems
  - Does not require complete access to the model, i.e., can attack ML as a service system
  - Extends to other modalities, e.g., natural language processing systems



Light projection attack demo

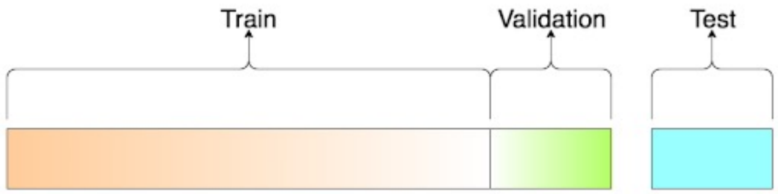
Attacking real-world Yolo-v5 detector running on Nvidia Xavier chip using MIPI camera feed (attacker does not need access to detection system)

- Make coffee cup invisible to the detector
- Fooling Yolo-v5 to incorrectly detect cup as a scissor



Existing validation approaches give false sense of security

- Common robustness evaluation practice is to train a system on a training data set, and then test it on another set



- This is insufficient to provide security guarantees as an attacker/nature can send inputs that differ from the test set

- Almost all the heuristic defenses have been "broken" soon after they were proposed

Defense	Accuracy
Buckman et al. (2018)	0%*
Ma et al. (2018)	5%
Guo et al. (2018)	0%*
Dhillon et al. (2018)	0%
Xie et al. (2018)	0%*
Song et al. (2018)	9%*

Is AI/ML useless for high-regret applications?

- Can we ever design deep neural networks (DNNs) that cannot be fooled with certain known unknown attacks or guarantee predictable behavior to achieve safe operation in many real-world applications?

- This might appear impossible given the following popular beliefs
  - Deep Learning is a black-box
  - No one knows why and how Deep Learning works
  - There are no guarantees with Deep Learning

Foolproof/Certified ML

Yet Another AI Snake Oil?

LLNL’s Foolproof AI: formal verification and provably robust design

Our **Blue AI team** has developed automated tools to make ML systems foolproof

- guarantee a self-driving car will always stop on a stop sign

ML Safety Toolbox

- Provable robustness analysis on any neural network structures (Verification)
- Differentiability and ease of use of our framework allow us to train foolproof ML (Design)

“Nothing is more useless than theory and guarantees that do not hold in practice”
-- Unknown

But how does it work?

Foolproof defense relies on our ability to "verify the robustness" of a given DNN

- Using formal methods to rigorously prove that certain properties hold

- Want to ensure: for a given input $\bar{x}_0$ and a given amount of noise δ, classification remains the same
- $P(\bar{x})$:
  - $\|\bar{x} - \bar{x}_0\|_{L_\infty} \leq \delta$
  - Equivalent to: $\bigwedge_i (-\delta \leq \bar{x}[i] - \bar{x}_0[i] \leq \delta)$
- $Q(\bar{y})$:
  - $\bigvee_i (\bar{y}[i_0] \leq \bar{y}[i])$, where $\bar{y}[i_0]$ is the desired label

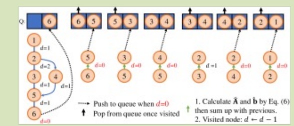
Our Magic Sauce: relaxation based formal verification

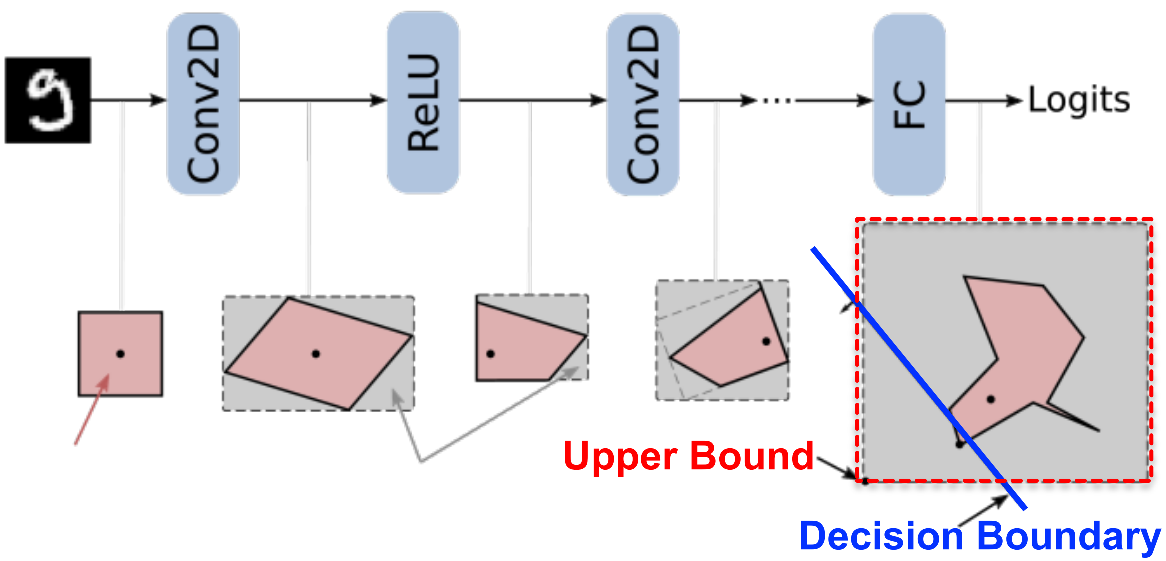
We employ linear relaxation techniques to compute provable linear bounds on DNN output

- obtain linear relaxations of any non-linear units
- “glue” these relaxations according to the network structure (or a compute graph)

ML Safety Toolbox

Statistics
Optimization
Software Eng.

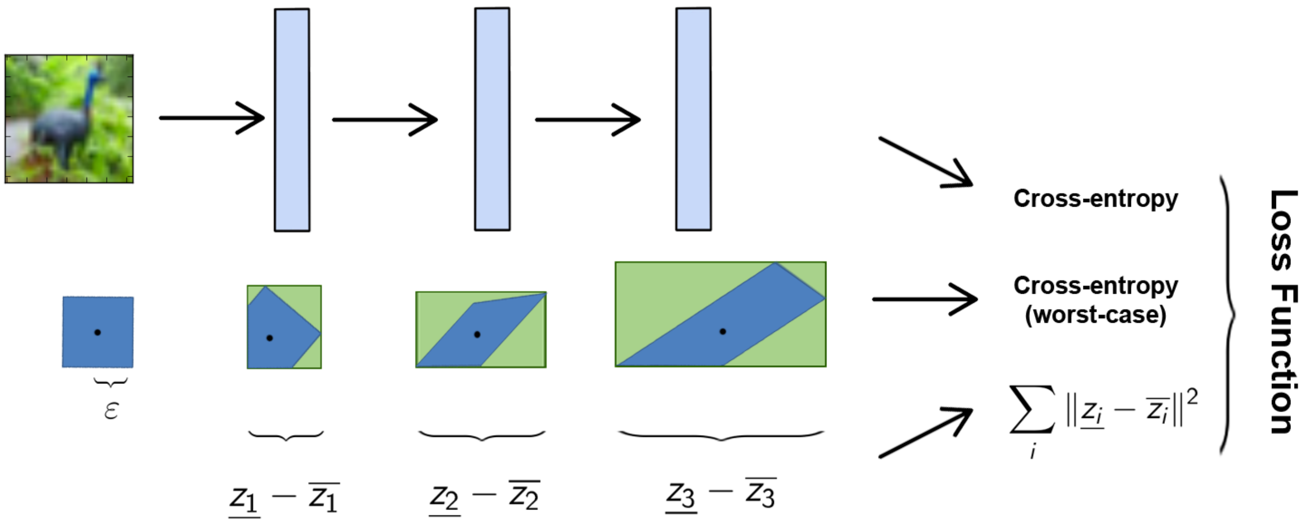




Foolproof AI by design

These bounds can be combined with training to design provably robust DNNs

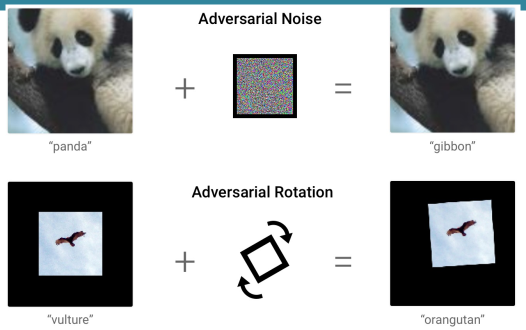
- ensure that the whole bounding box is classified correctly



What are we able to achieve?

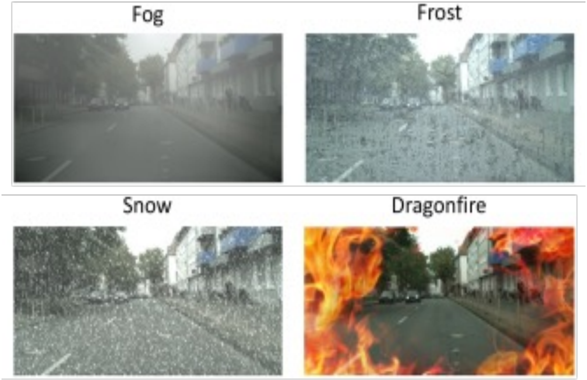
Foolproof for adversarial shifts

- Imperceptible Perturbations
- Geometric Perturbations
- Any shift that can be modelled (e.g., simple natural shifts, logic tables)



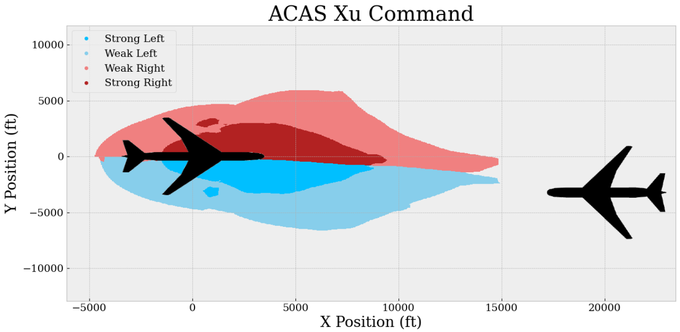
Foolproof for common corruptions

- Complex natural shifts
- Shifts in scientific domains



Provably enforcing certain application specifications

- Unmanned airborne collision avoidance system (ACAS-Xu)



We can achieve certified accuracy

- Certified robustness on complicated networks that could not be supported by prior work
- Certified defense on ImageNet where previous approaches could not scale

Robust Vision Models with l_inf attack

Dataset	DenseNet	W-ResNet	ResNeXt
CIFAR10	32.43%	32.23%	31.75%
ImageNet	14.56%	15.86%	13.05%

Robust NLP Models with substitution attack

Model	2-word	4-word	6-word
LSTM	23.4%	23.4%	23.4%
Transformer	22.6%	22.6%	22.6%

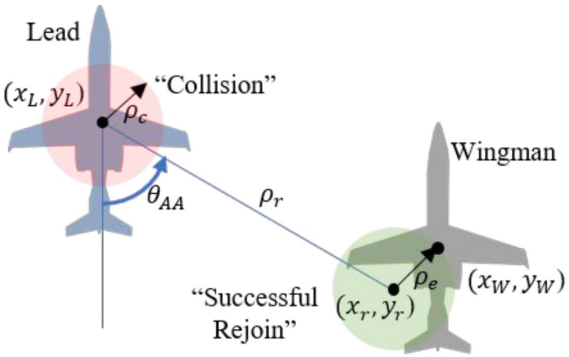
These numbers imply that an adversary cannot fool these many test samples regardless of the amount of compute it throws at any adversarial example generation algorithm

Cutting-edge science to real-world impact on safety-critical applications

- Winner of International Verification of Neural Networks Competition (VNN-COMP 2022) α,β-CROWN is built upon our AutoLiRPA technique

- The goal of the competition is to compare neural network verification methods, in terms of scalability and runtime speed
  - standard formats (ONNX for NNs and VNNLIB for specifications), hardware (AWS)

- In addition, to verifying standard vision benchmarks (CIFAR classifiers), α,β-CROWN performed the best on
  - ACAS-XU airborne collision avoidance benchmark
  - AFRL ACT3's SafeRL benchmark for aircraft rejoin



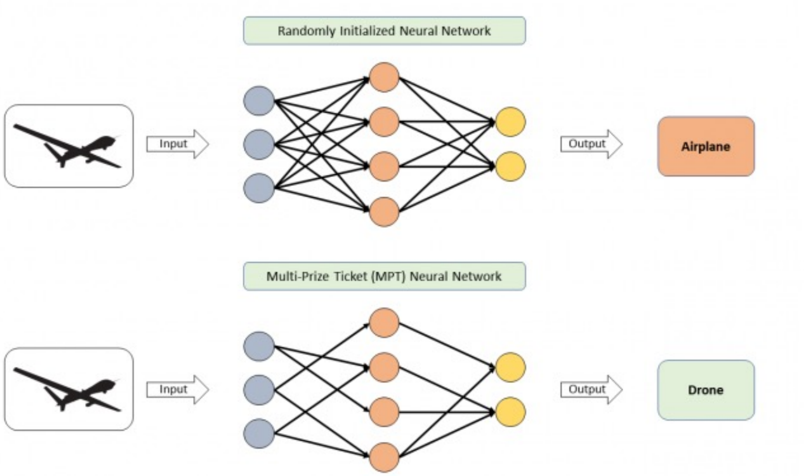
Existing DNNs are not suitable for real-time and resource-limited applications

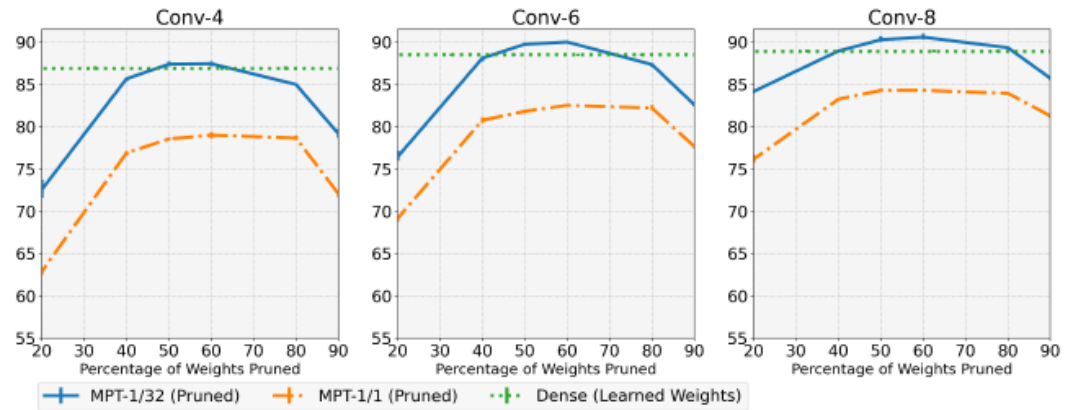
- First positive result on designing CARDs
  - **C**ompact – small in size (reduction from 1gb to <1mb) and latency (reduction from 100ms to <1ms)
  - **A**ccurate – state-of-the-art accuracy
  - **R**obust – graceful degradation
  - **D**eep Neural Nets

- Our tools are not image specific and apply to text/tabular modality

A new paradigm for learning efficient DNNs

We proposed a new paradigm for learning neural networks – instead of iteratively weight-training, we simply prune and binarize weights (Multi-Prize Tickets (MPT))





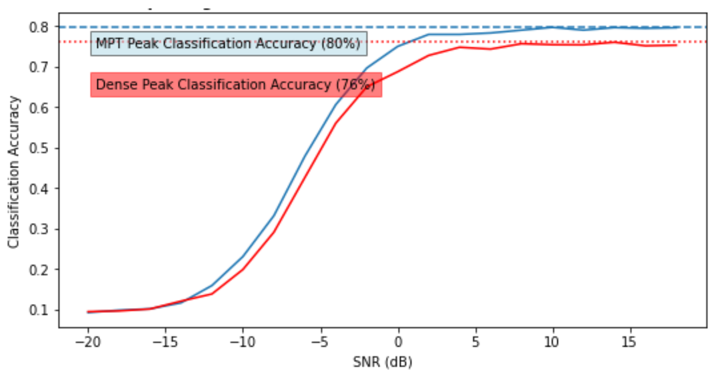
- MPTs result in ~32× memory saving and ~58× computation saving
- Top model in RobustBench leaderboard

Diffenderfer, J., & Kailkhura, B. "Multi-Prize Lottery Ticket Hypothesis: Finding Accurate Binary Neural Networks by Pruning A Randomly Weighted Network". ICLR 2021.

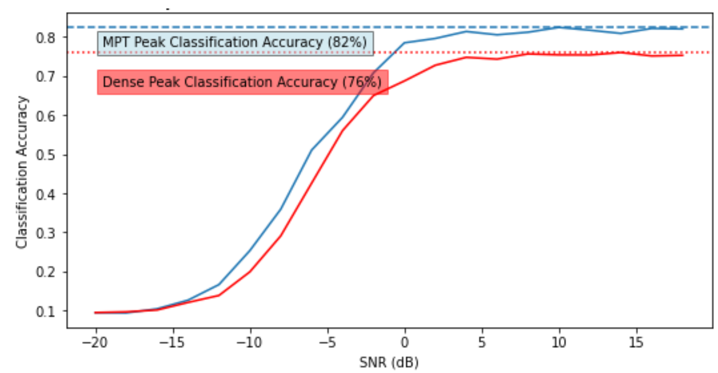
We have developed a RF-ML system that is ~500x smaller and ~50x faster

- Application to radio frequency ML system
  - Develop a signature detection and classification system for Army tactical vehicles, to reduce cognitive burden on Army signals analysts

500x smaller, 50x faster



500x smaller, 2x faster



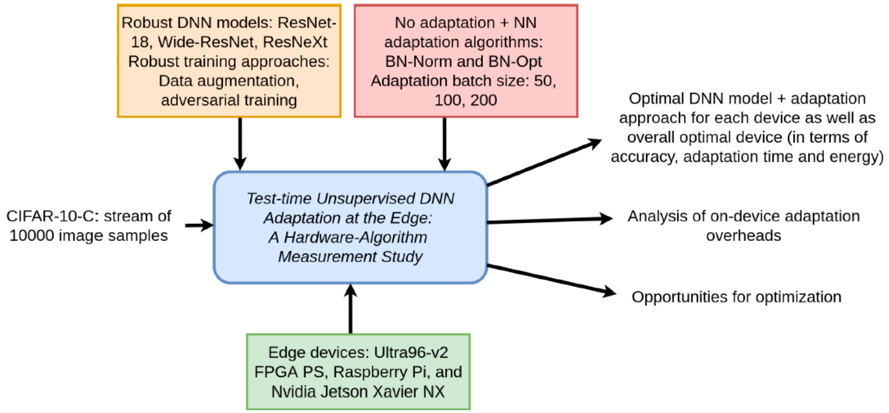
6% better top accuracy

Better robustness at all SNRs (20db dense = 0db MPT)

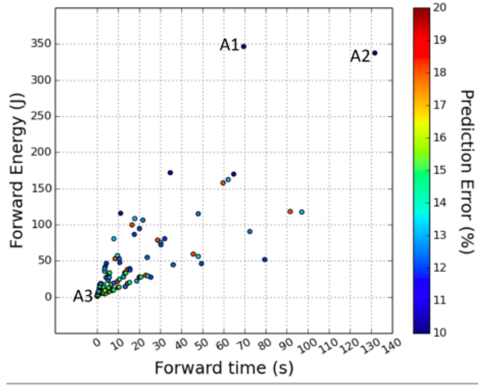
- Developing low-power hardware AI chip for real-world demonstration (100x energy gains)

Characterization on edge devices

- We characterize the test-time adaptation performance of standard neural nets on corrupted CIFAR-10 at edge devices
  - FPGA, Raspberry-Pi, and Nvidia Xavier NX



- Our characterization provided some very interesting results
  - Approach that only updates the normalization parameters with Wide-ResNet, running on Xavier GPU, to be overall effective in terms of balancing multiple cost metrics
  - However, the adaptation overhead is extremely high (around 213 ms)



A1: RXT-AM-200 + BN-Opt + NX-CPU
A2: RXT-AM-200 + BN-Opt + RPi
A3: WRN-AM-50 + BN-Norm + NX-GPU

Fig. 12: Overall results with all the points from Figs. 5, 8, 11. A1/A2: when accuracy is the only priority, A1 shows the lowest runtime and A2 the lowest energy (i.e. among all points with 10.15% error). A3: optimal point when all three costs are equally important (0.31s, 2.96J, 15.21%).

- Our results strongly motivate the need for algorithm-hardware co-design for efficient on-device DNN adaptation

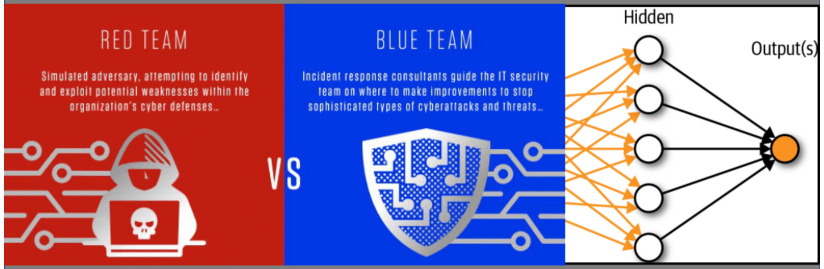
Takeaways from this talk

- Deep learning in real-world systems is probably here to stay
- It is possible to verify important properties of DNNs and design Foolproof AI
- It is possible to achieve efficiency and performance simultaneously

Ongoing efforts for ensuring that AI systems in the real world do the "right thing"

- Broadening the scope of the adversary
- Efficient training and inference schemes for LLMs/VLMs
- Co-design for efficient AI

LLNL's Mission-Critical AI



We can develop predictable, assured, and efficient ML systems

- ML verification tools to formally prove that models are robust to a range of attacks
- Assured design tools to train ML models that are accurate as well as specification consistent
- ML compression tools to design ultra compact neural nets and low power AI chips
- We can support a range of mission-critical applications (vision, NLP, tabular, etc.)

CASC

Center for Applied Scientific Computing

Lawrence Livermore National Laboratory