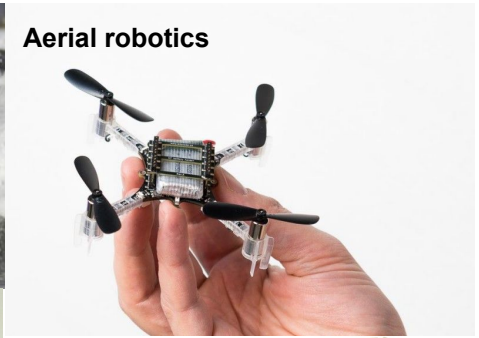
TinyML has a Security Problem - An Adversarial Perturbation Perspective

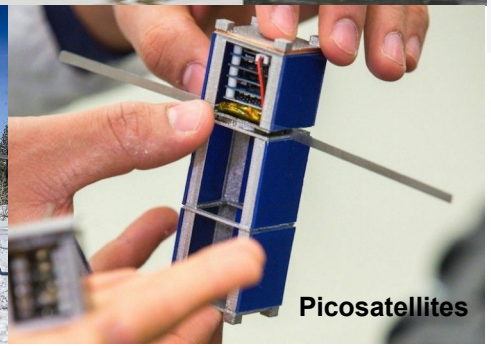
Swapnil Sayan Saha, Khushbu Pahwa, and Basheer Ammar

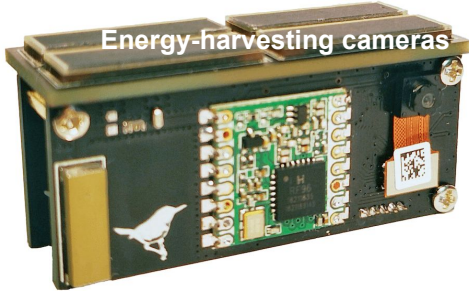
ECE 209AS Winter 2022 Final Project

UCLA Samueli School of Engineering

Motivation: AI-enabled IoT



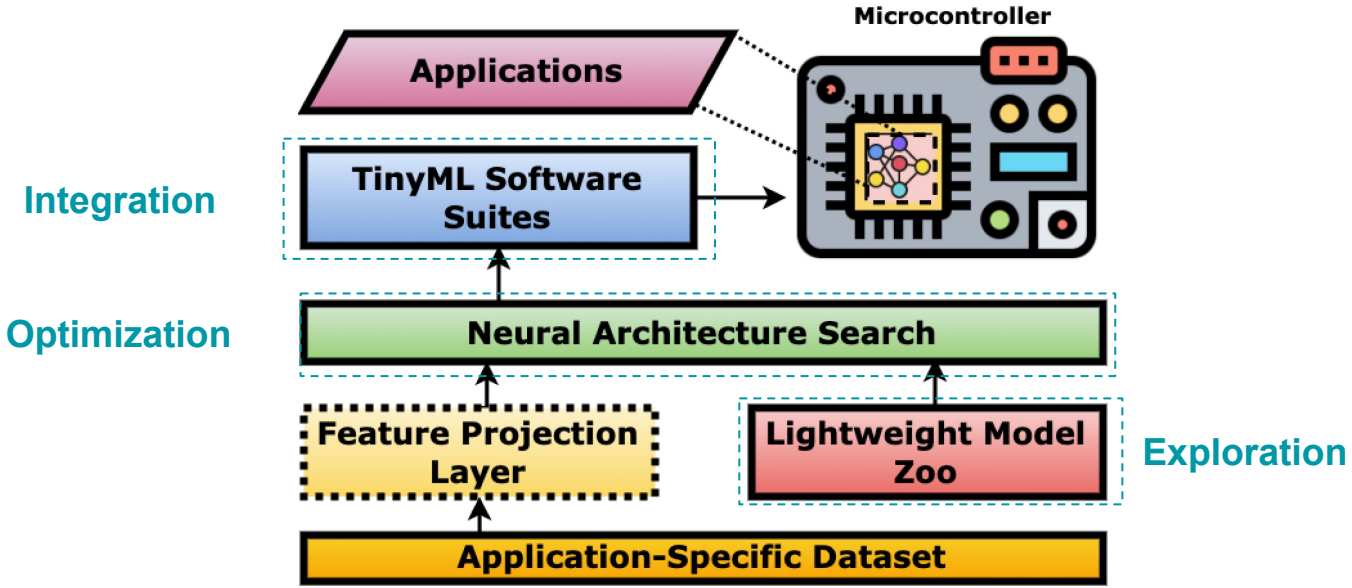




AI-enabled IoT - key to making “**complex inferences”** for “**time-critical”** and “**remote”** applications from “**unstructured data”**.

Swapnil Sayan Saha, Sandeep Singh Sandha, and Mani Srivastava, “Machine Learning for Microcontroller-Class Hardware – A Review”, in I*EEE Sensors Journal*, 2022. (under review).
Sanchez-Iborra, Ramon, and Antonio F. Skarmeta. "Tinyml-enabled frugal smart objects: Challenges and opportunities." *IEEE Circuits and Systems Magazine* 20.3 (2020): 4-18.
Warden, Pete, and Daniel Situnayake. *TinyML*. O'Reilly Media, Incorporated, 2019

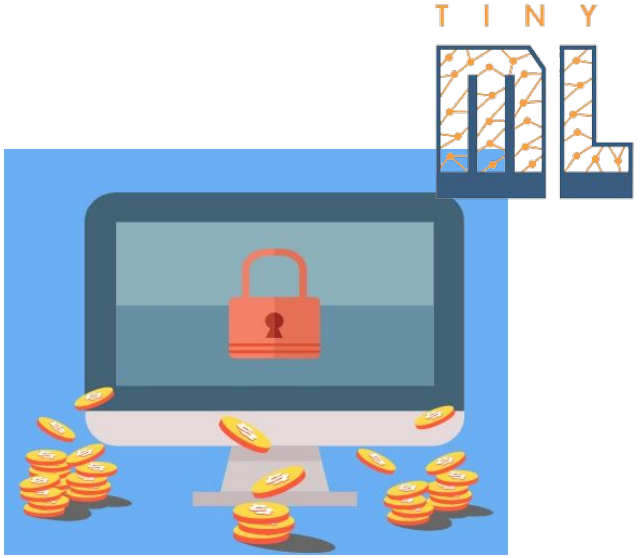
The TinyML Workflow



First-generation efforts (dubbed TinyML) focused on the **exploration**, **optimization** and **integration** of simple neural networks to **low-end** IoT devices.

Swapnil Sayan Saha, Sandeep Singh Sandha, and Mani Srivastava, “Machine Learning for Microcontroller-Class Hardware – A Review”, in I*EEE Sensors Journal*, 2022. (under review).
Warden, Pete, and Daniel Situnayake. *TinyML*. O'Reilly Media, Incorporated, 2019.
C. Banbury, V. J. Reddi, P. Torelli, J. Holleman, N. Jeffries, C. Kiraly, P. Montino, D. Kanter, S. Ahmed, D. Pau, et al., “Mlperf tiny benchmark,” *Advances in Neural Information Processing Systems,* 2021.
Ray, Partha Pratim. "A review on TinyML: State-of-the-art and prospects." *Journal of King Saud University-Computer and Information Sciences* (2021).

Problem: No Security Analysis in the Workflow



No attack surface analysis at various CPS layers in the workflow.

Lack of quantification of security costs of lightweight models.

Swapnil Sayan Saha, Sandeep Singh Sandha, and Mani Srivastava, “Machine Learning for Microcontroller-Class Hardware – A Review”, in I*EEE Sensors Journal*, 2022. (under review).

Contributions

Quantify how TinyML models are less robust to adversarial attacks over large models.

Provide an automated and efficient AutoML solution to generate adversarially robust TinyML models within hardware constraints.

Attack Model

Goal: Covert and passive perturbation of inputs to cause the model provide erroneous outputs.

No white-box or black-box access to the model.

Applications: Image recognition, audio keyword spotting and wearable human activity detection.

Universal adversarial perturbation^ via well-known techniques*.

^Moosavi-Dezfooli, Seyed-Mohsen, et al. "Universal adversarial perturbations." Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.
^Zhou, Wen, et al. "Transferable adversarial perturbations." Proceedings of the European Conference on Computer Vision (ECCV). 2018.
*Zhao, Yue, et al. "Seeing isn't believing: Towards more robust adversarial attack against real world object detectors." Proceedings of the 2019 ACM SIGSAC Conf. on Computer and Communications Security. 2019.
*Eykholt, Kevin, et al. "Robust physical-world attacks on deep learning visual classification." Proceedings of the IEEE conference on computer vision and pattern recognition. 2018.
*Zhang, Guoming, et al. "Dolphinattack: Inaudible voice commands." Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security. 2017.
*Li, Juncheng, et al. "Adversarial music: Real world audio adversary against wake-word detection system." Advances in Neural Information Processing Systems 32 (2019).
*Sugawara, Takeshi, et al. "Light Commands:{Laser-Based} Audio Injection Attacks on {Voice-Controllable} Systems." 29th USENIX Security Symposium (USENIX Security 20). 2020.
*El-Rewini, Zeinab, et al. "Cybersecurity attacks in vehicular sensors." IEEE Sensors Journal 20.22 (2020): 13752-13767.

Candidate Models

Application	Model Type	Model	Test Accuracy (%)	Parameters (M)	Features	NAS@	Transfer Learning
Image Recognition	Large	EfficientNetB0 [30]	93.2	4.07	✗	✓	✓*
		EfficientNetB4 [30]	93.5	17.70	✗	✓	✓*
		EfficientNetv2B0 [31]	96.7	5.93	✗	✓	✓*
		EfficientNetv2B3 [31]	97.0	12.95	✗	✓	✓*
		ResNet50 [32]	89.7	23.62	✗	✗	✓*
		VGG19 [33]	92.3	20.03	✗	✗	✓*
	TinyML	ResNet8 [29]	87.1	0.079	✗	✗	✗
		MCUNet DS-CNN (320-1)` [34]	87.7	0.57	✗	✓	✓*
		MCUNet DS-CNN (256-1)` [34]	87.5	0.56	✗	✓	✓*
Audio Keyword Spotting	Large	Attention RNN [35]	93.9	1.29	✓^	✗	✗
		CNN [36]	82.4	0.095	✓^	✗	✗
		GRU [36]	92.2	0.50	✓^	✗	✗
		LSTM [36]	92.9	1.03	✓^	✗	✗
	TinyML	DS-CNN [29]	92.2	0.025	✓^	✗	✗
		TCN [37][38]	76.0	0.019	✓^	✗	✗
Human Activity Recognition	Large	CNN-LSTM [39]	99.7	1.74	✗	✗	✗
		CNN [39]	99.0	3.03	✗	✗	✗
		LSTM [39]	97.3	0.20	✗	✗	✗
	TinyML	Bonsai [40]	72.6	0.00063	✓	✗	✗
		ProtoNN [41]	72.0	0.00062	✓	✗	✗
		TCN [37][38]	93.0	0.00106	✗	✗	✗
		FastRNN [42]	95.0	0.00015	✗	✗	✗
		FastGRNN [42]	98.6	0.00030	✗	✗	✗

* Transfer learning from ImageNet-1000 to CIFAR-10; ^ Operates on log-Mel spectrograms; @ Optimized via AutoML frameworks; ` Models targeted towards two different microcontrollers.

13 large models, 10 TinyML models.

3 representative applications and datasets.

Diversity in model architecture.

Attacks and Metrics

Attack 1: FGSM

$$x_p = x + \epsilon \cdot \text{sign}(\nabla_x J(f, x, y))$$

Attack 2: PGD

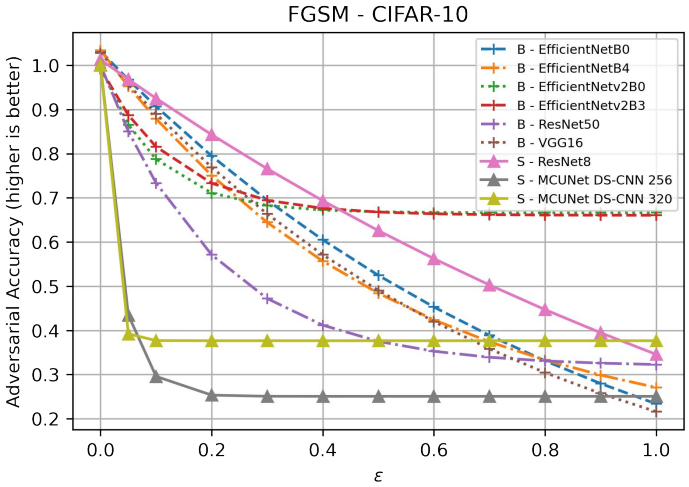
$$x_p^{t+1} = \text{clip}_\epsilon(x^t + \alpha \cdot \text{sign}(\nabla_x J(f, x^t, y)))$$

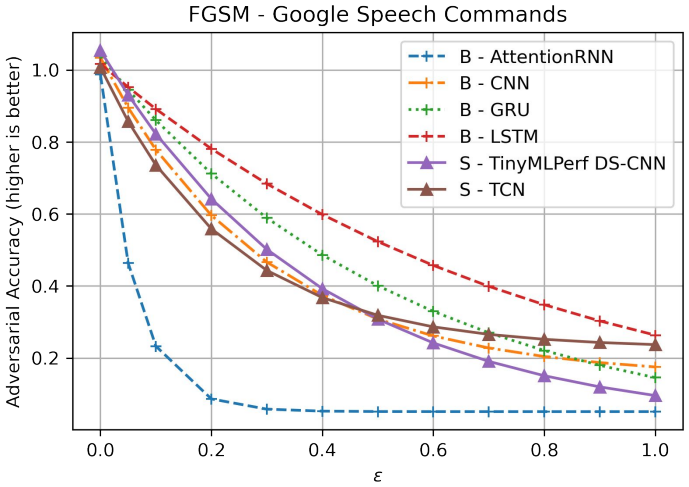
Metric: Adversarial Accuracy

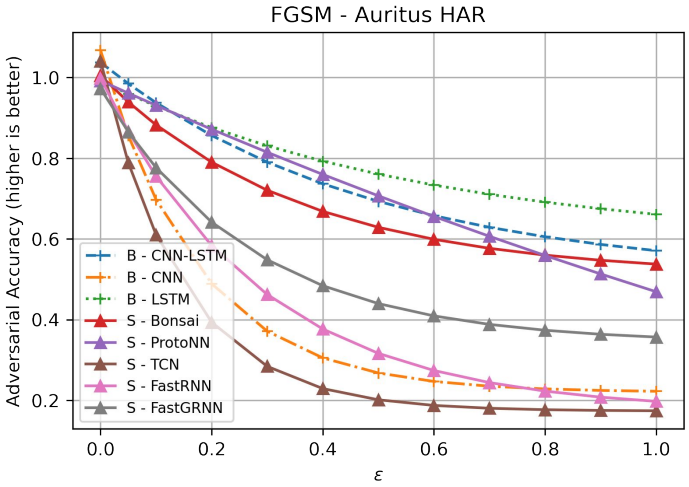
$$\frac{1}{N}\sum_{i=0}^{N} q_i, \qquad q_i = \begin{cases} 1, & \text{if} \quad y_{\text{pred}}^{x_i} = y_{\text{pred}}^{x_{i,p}} \\ 0, & \text{otherwise} \end{cases}$$

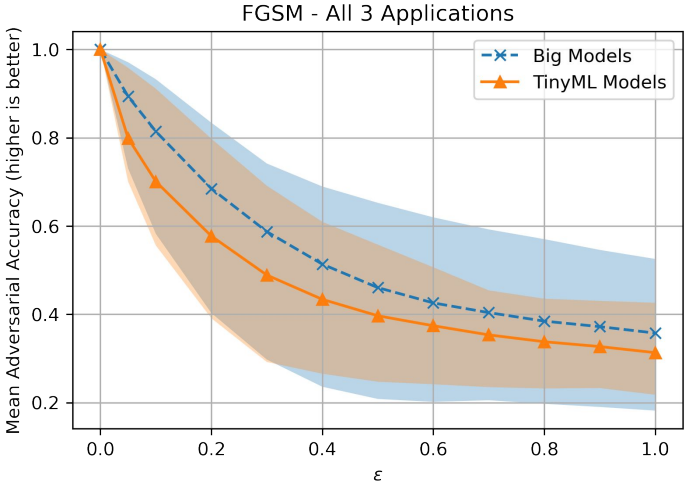
Goodfellow, Ian J., Jonathon Shlens, and Christian Szegedy. "Explaining and harnessing adversarial examples." *International Conference on Learning Representations* (2015).
Kurakin, Alexey, Ian Goodfellow, and Samy Bengio. "Adversarial machine learning at scale." *International Conference on Learning Representations* (2017).

Results of FGSM Attack

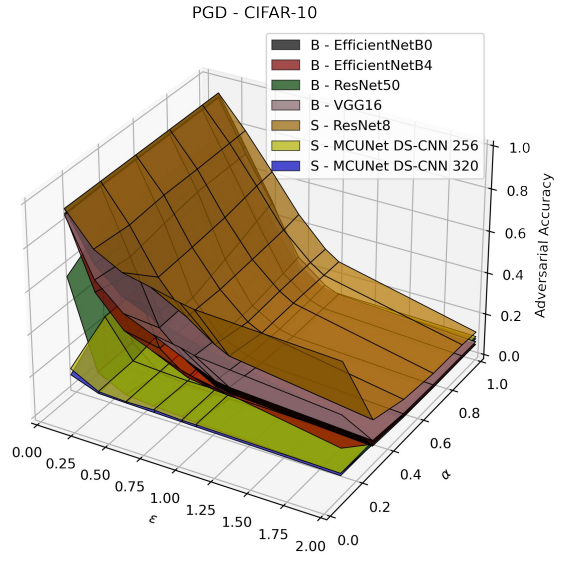


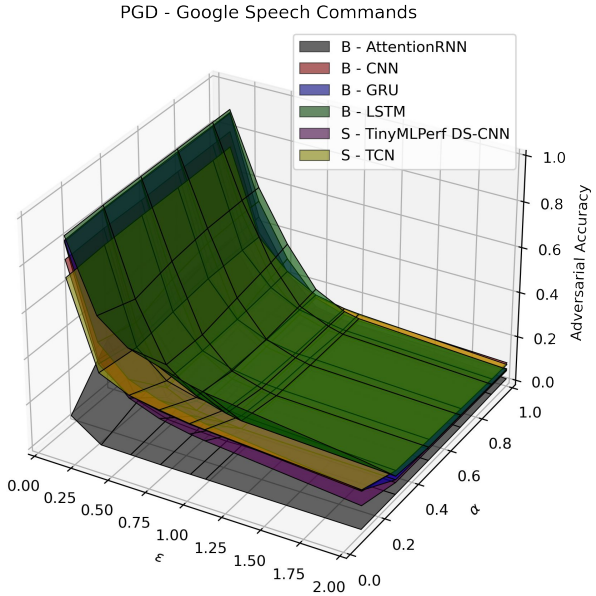


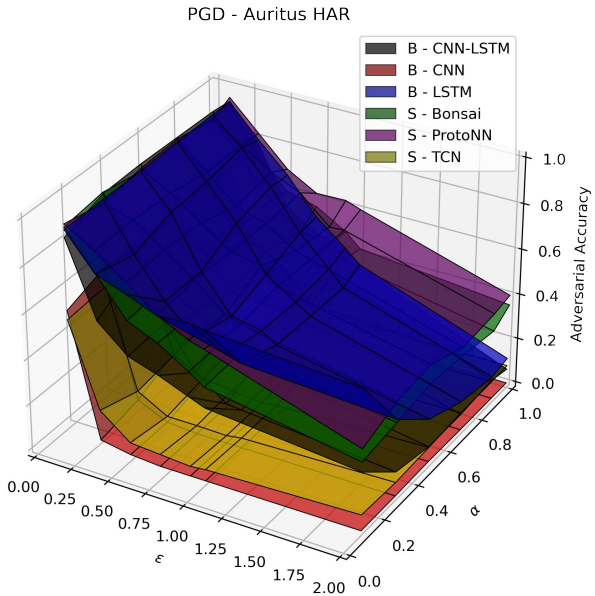




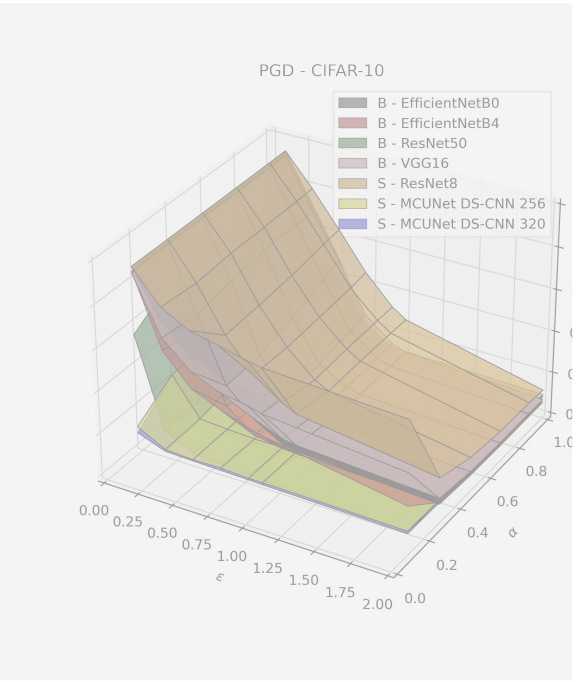
Results of PGD Attack

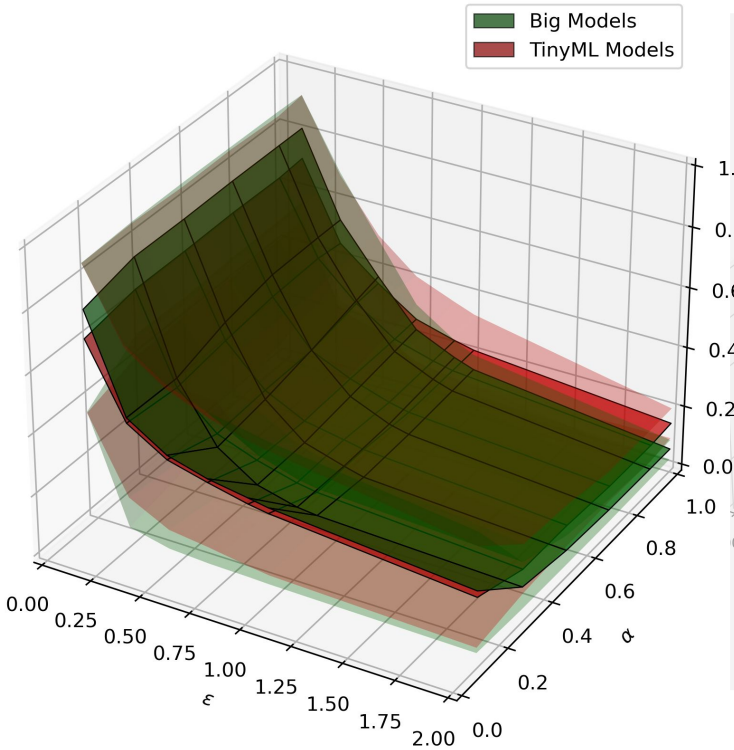


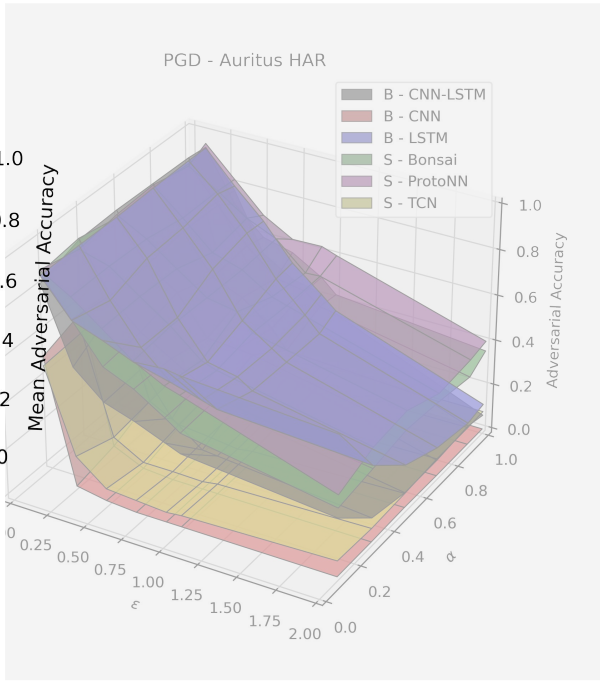




Results of PGD Attack







Solution: NAS with Adversarial Robustness

$$f_{\text{opt}} = \lambda_1 f_{\text{error}}(\Omega) + \lambda_2 f_{\text{flash}}(\Omega) + \lambda_3 f_{\text{SRAM}}(\Omega) + \lambda_4 f_{\text{latency}}(\Omega) + \lambda_5 f_{\text{adversarial error}}(\Omega)$$

$$f_{\text{error}}(\Omega) = \mathcal{L}_{\text{validation}}(\Omega), \Omega = \{\{V, E\}, w, \theta, v\}$$

$$f_{\text{flash}}(\Omega) = \begin{cases} -\frac{||h_{\text{FB}}(w,\{V,E\})||_0}{\text{flash}_{\text{max}}} \vee -\frac{\text{HIL information}}{\text{flash}_{\text{max}}} \\ \infty, f_{\text{flash}}(\Omega) > \text{flash}_{\text{max}} \end{cases}$$

$$f_{\text{SRAM}}(\Omega) = \begin{cases} -\frac{\max_{l \in [1,L]} \{||x_l||_0 + ||a_l||_0\}}{\text{SRAM}_{\text{max}}} \vee -\frac{\text{HIL information}}{\text{SRAM}_{\text{max}}} \\ \infty, f_{\text{SRAM}}(\Omega) > \text{SRAM}_{\text{max}} \end{cases}$$

$$f_{\text{latency}}(\Omega) = \frac{\text{FLOPS}}{\text{FLOPS}_{\text{target FLOPS}}} \vee \frac{\text{HIL information}}{\text{Latency}_{\text{target latency}}}$$

$$f_{\text{adversarial error}}(\Omega) = 1 - \frac{1}{N} \sum_{i=0}^{N} q_i, \quad q_i = \begin{cases} 1, & \text{if} \quad y_{\text{pred}}^{x_{i,\text{validation}}} = y_{\text{pred}}^{x_{i,\text{validation},p}} \\ 0, & \text{otherwise} \end{cases}$$

$$\hat{f}(\Omega) \sim \mathcal{GP}(\mu(\Omega), k(\Omega, \Omega'))$$

$$\Omega_t = \arg\max_{\Omega} (\mu_{t-1}(\Omega) + \beta^{0.5} \sigma_{t-1}(\Omega))$$

- Yields adversarially robust models cheaply.
- Handles categorical variables.
- Usage of models beyond simple CNN.
- Not affected by discontinuity in loss contour.
- Converges within 10-50 epochs.

Sandeep Singh Sandha, Mohit Aggarwal, Swapnil Sayan Saha, and Mani Srivastava. "Enabling Hyperparameter Tuning of Machine Learning Classifiers in Production" *Third IEEE International Conference on Cognitive Machine Intelligence*. IEEE, 2021. 1-10.

Results of Robust NAS

Model	Test Accuracy (%)	Adversarial Accuracy	Model Size (kB)	FLOPS (M)
TCN (STM32F746ZG*, handcrafted)	93	18	68.8	12.57
TCN (NAS, STM32F746ZG, no adversarial term)	90	20	**44.4**	**8.52**
TCN (NAS, STM32F746ZG, with adversarial term)	**97**	**44**	101.4	23.2
TCN (NAS, STM32F446RE*, no adversarial term)	93	14	**54.93**	**8.79**
TCN (NAS, STM32F446RE, with adversarial term)	**93**	**27**	68.95	15.87
TCN (NAS, STM32L476RG*, no adversarial term)	90	14.3	**38.7**	**5.37**
TCN (NAS, STM32L476RG, with adversarial term)	**95**	**36.3**	96.97	21.57
ProtoNN (handcrafted)	**72**	**78**	27.8	-
ProtoNN (NAS, no adversarial term)	71	72	**1.16**	-
ProtoNN (NAS, with adversarial term)	70	76	20.4	-
Bonsai (handcrafted)	73	19.6	14.9	-
Bonsai (NAS, no adversarial term)	71	18.6	1.5	-
Bonsai (NAS, with adversarial term)	72	**20.6**	**1.5**	-

Conclusion and Future Work

We empirically showed security issues in TinyML from an adversarial perturbation standpoint.

We provide an inexpensive AutoML solution to the problem.

Starting point to make TinyML models more backward compatible with upstream models without additional cost

Need to provide a generalized and universal pipeline in the TinyML workflow for other attacks.