Institute for the Wireless Internet of Things
at Northeastern University

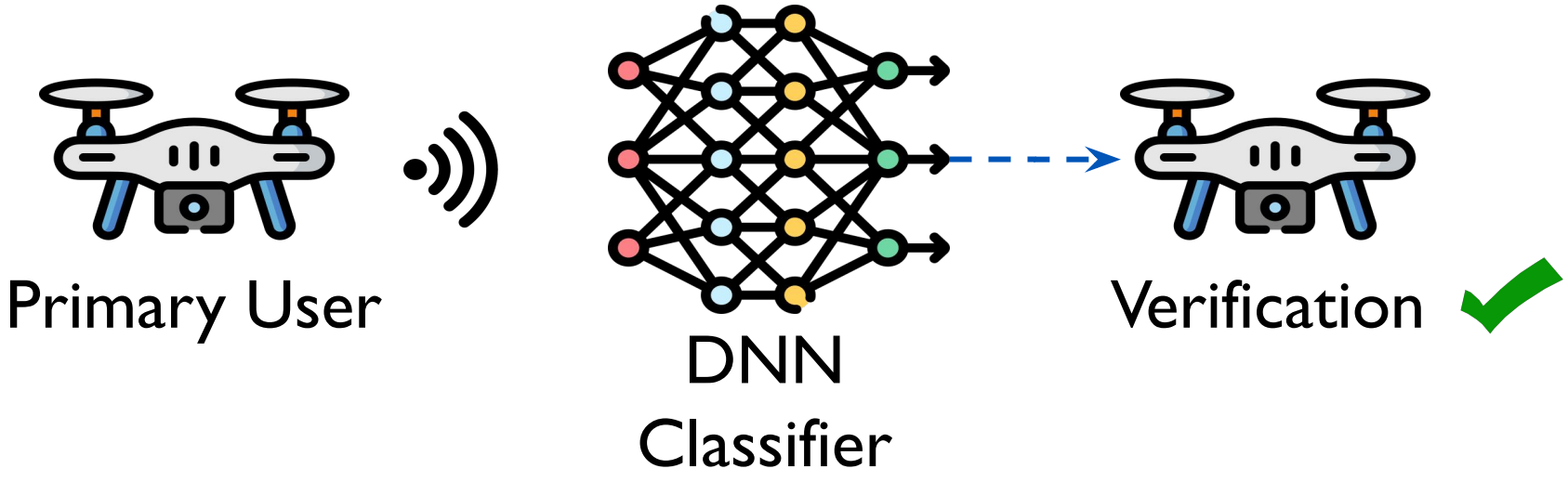
HyperAdv: Dynamic Defense Against Adversarial Radio Frequency Machine Learning Systems

Milin Zhang, Michael De Lucia, Ananthram Swami,
Jonathan Ashdown, Kurt Turck and Francesco Restuccia

Presenting: **Milin Zhang**
Northeastern University, United States
Email: frestuc@northeastern.edu
Website: **http://mentislab.info**

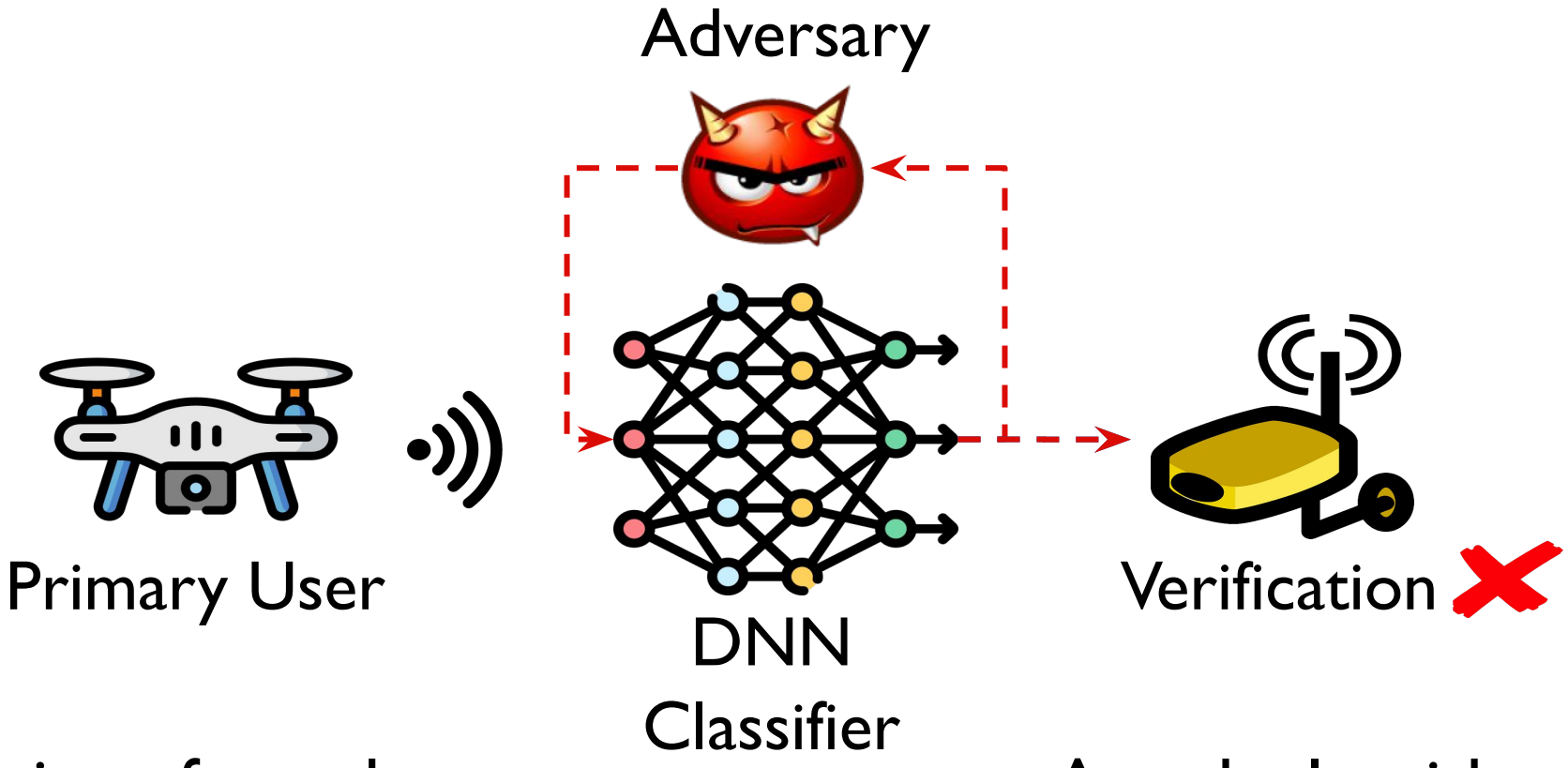
Adversarial Machine Learning

Institute for the Wireless Internet of Things at Northeastern



Adversarial Machine Learning

Institute for the Wireless Internet of Things at Northeastern



Objective of attacker:

$$f(x+\delta) = y' \quad y' \neq y$$

Attack algorithm [1]:

$$x_{t+1}^{adv} = x_t^{adv} + \alpha \cdot \frac{\nabla \mathcal{L}(f(x_t^{adv}), y)}{||\nabla \mathcal{L}(f(x_t^{adv}), y)||_p}$$

[1] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards Deep Learning Models Resistant to Adversarial Attacks," in Proceedings of International Conference on Learning Representations (ICLR), 2018.

Static Adversarial Defense

Min-max optimization in adversarial training (AT) [1]:

$$\min_{f(\cdot)} \mathbb{E}\{\max_{\delta} \mathcal{L}(f(x+\delta), y)\}$$

Outer loop: optimize DNN with adversarial samples

Inner loop: find adversarial samples

	Natural	Adversarial
Benign	**92.7%**	79.4% (-13.3%)
Adversarial	0.8%	**43.7%** (+41.9%)

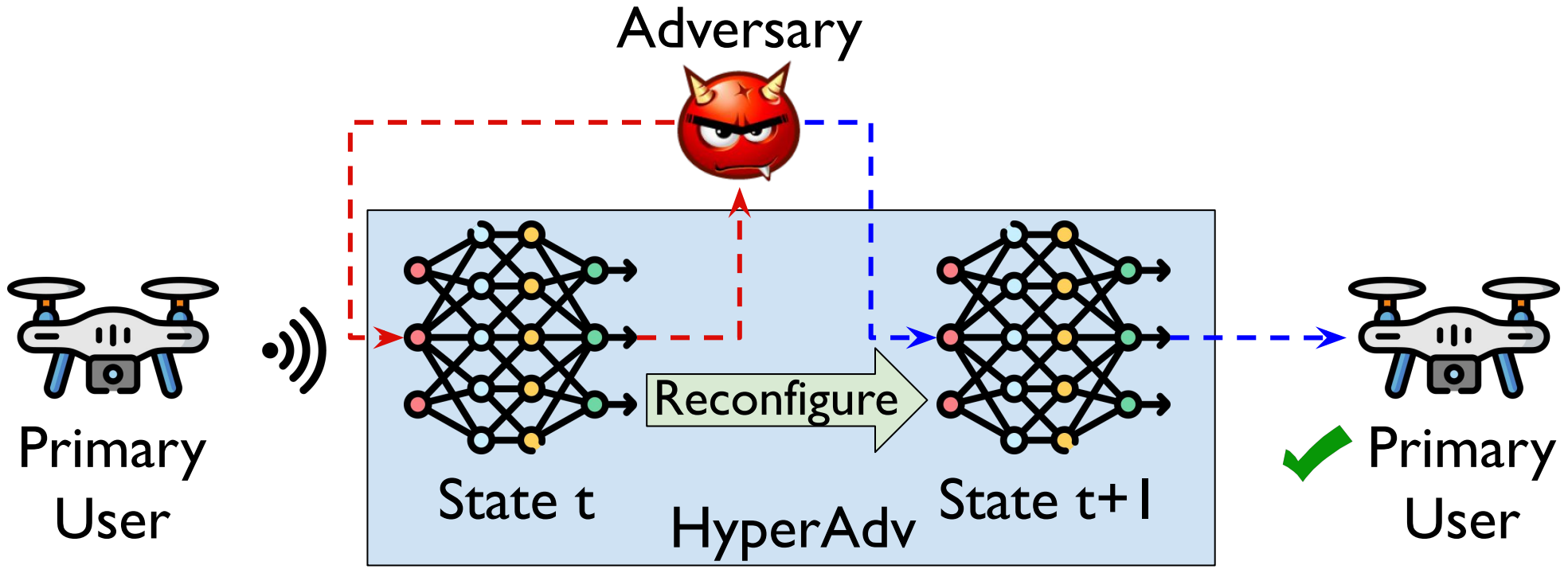
Natural Training (NT) and AT on CIFAR 10 (data reported in original paper [1])

Adversarial training sacrifice performance on clean data

[1] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards Deep Learning Models Resistant to Adversarial Attacks," in Proceedings of International Conference on Learning Representations (ICLR), 2018.

Dynamic Adversarial Defense: HyperAdv

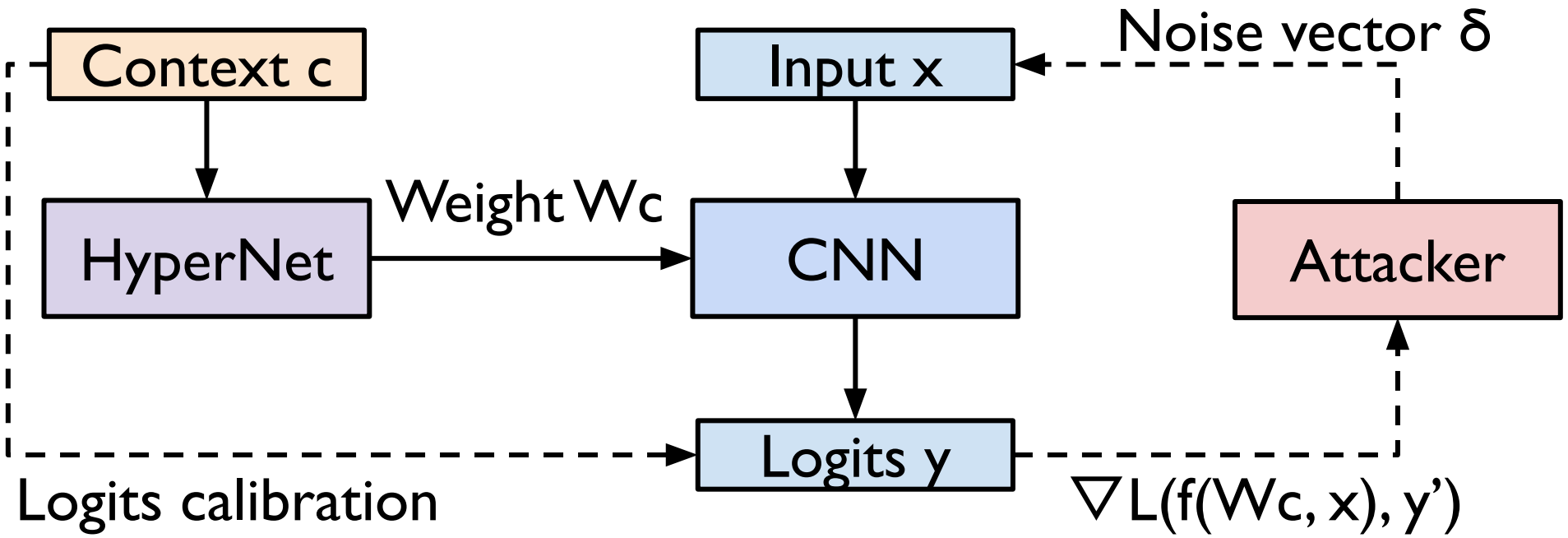
Institute for the Wireless Internet of Things at Northeastern



Key idea: dynamically change the DNN at each inference step so that the adversarial information at step t is not effective at step t+1

HyperAdv Framework

Institute for the Wireless Internet of Things at Northeastern



HyperNet [1]: a neural network to generate weight Wc for CNN classifier dynamically

CNN: a target network to perform RF signal classification

Attacker: Adversary to create noise δ based on the gradient $\nabla L(f(Wc, x), y')$

Context: Random vector to trigger the weight generation as well as calibrate the output y to y'

[1] D. Ha, A. M. Dai, and Q.V. Le, "HyperNetworks," in International Conference on Learning Representations (ICLR), 2017.

Ensemble Learning Strategy

Institute for the Wireless Internet of Things at Northeastern

Training:

1. a set of context $C_i = [C_i^1, C_i^2]$ is randomly sampled from $U(-1, 1)$
2. HyperNet $H(\cdot) : C_i \mapsto W_c^i$ generate weights for target CNN
3. CNN $f_{W_c^i}(\cdot) : x \mapsto y$ maps input to label space
4. logit calibration $y_i' = C_i^1 * y + C_i^2$ to map output to diversified space
5. Ensemble learning

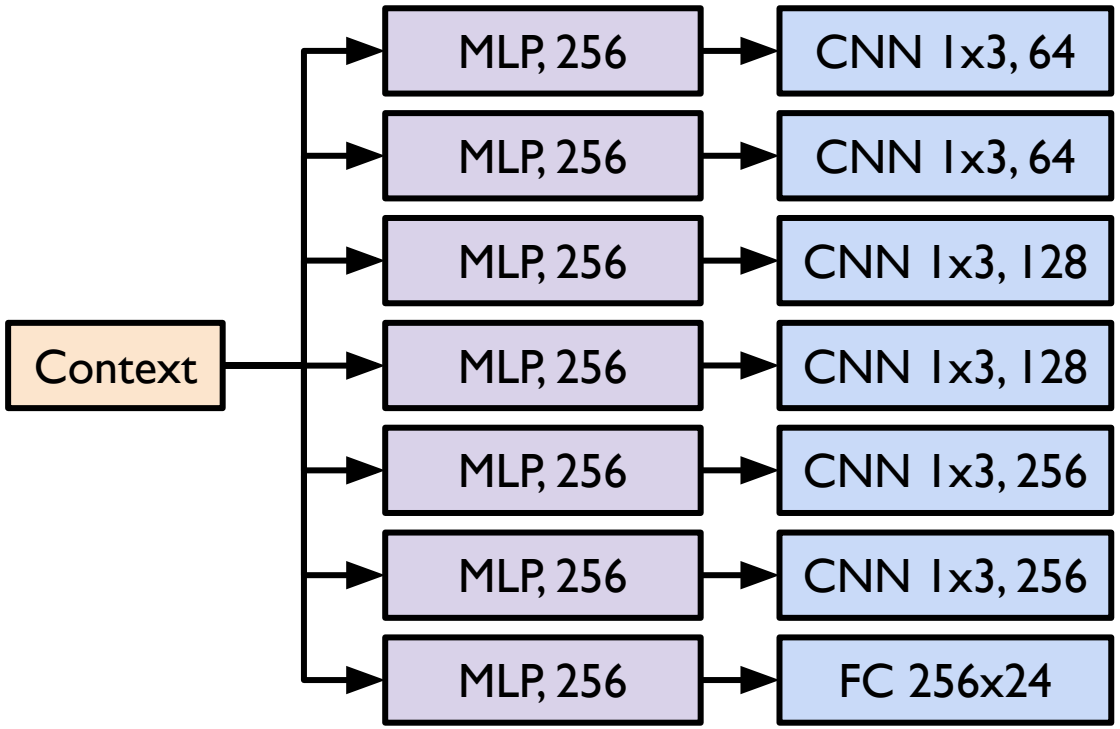
$$\mathcal{L} = \frac{1}{n} \sum_i \mathcal{L}_i(f_{W_c^i}(x), y_i')$$

can combine other defensive training such as AT and TRADES [1] to further improve robustness

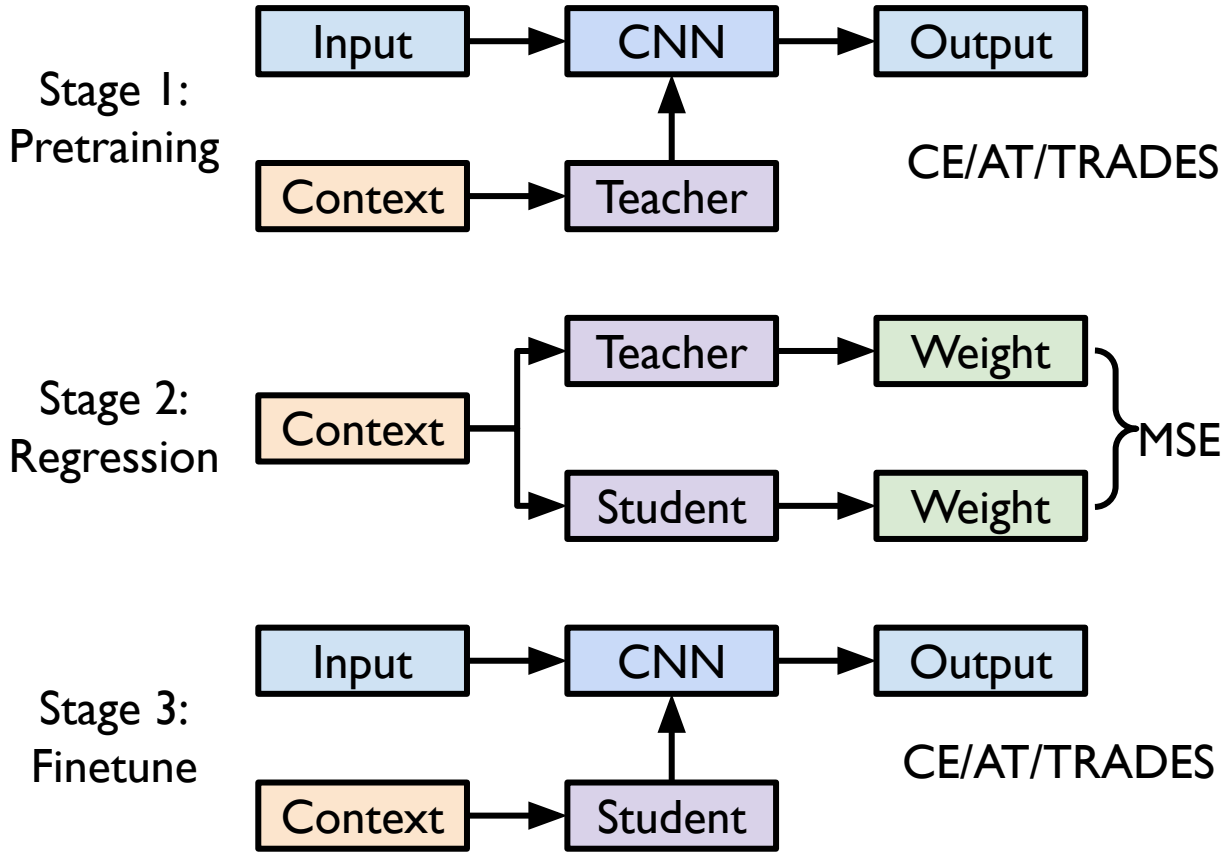
[1] H. Zhang, Y. Yu, J. Jiao, E. Xing, L. El Ghaoui, and M. Jordan, "Theoretically Principled Trade-off between Robustness and Accuracy," in Proceedings of International Conference on Machine Learning (ICML), pp. 7472–7482, PMLR, 2019.

Teacher-Student Design

Institute for the Wireless Internet of Things at Northeastern



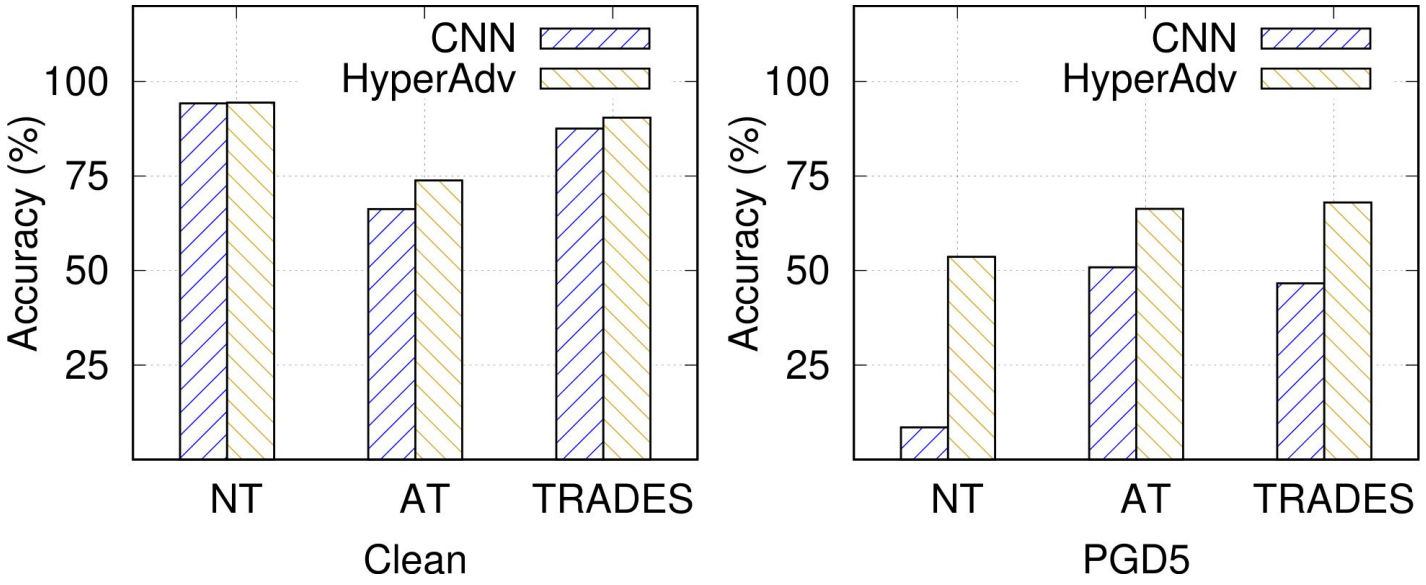
The size of Hypernetwork is **~M times larger** (e.g. 256) than a single target CNN



Effectively reduce the size by **~32 times**

Experimental Results

Institute for the Wireless Internet of Things at Northeastern



Robustness-Performance Trade-off:

AT and TRADES reduce accuracy on clean data by 27.92% and 6.65%

Robustness:

48% improvement compared to naturally trained DNN;

16% compared to AT;

22% compared to TRADES;

Accuracy:

Increase by 1%, 8% and 3% respectively.

Experimental Results

Institute for the Wireless Internet of Things at Northeastern

TABLE I: Accuracy of HyperAdv and its static counterparts

	NT		AT		TRADES	
	Clean	PGD	Clean	PGD	Clean	PGD
CNN	94.20	3.72	66.28	50.42	87.55	43.16
HyperAdv-R	95.20	**45.02**	74.01	**63.98**	90.37	**60.92**
HyperAdv-S	95.11	23.70	72.74	50.10	91.27	43.08
HyperAdv-E	**96.34**	27.88	**78.36**	61.34	**92.79**	49.50

Ablation Study:

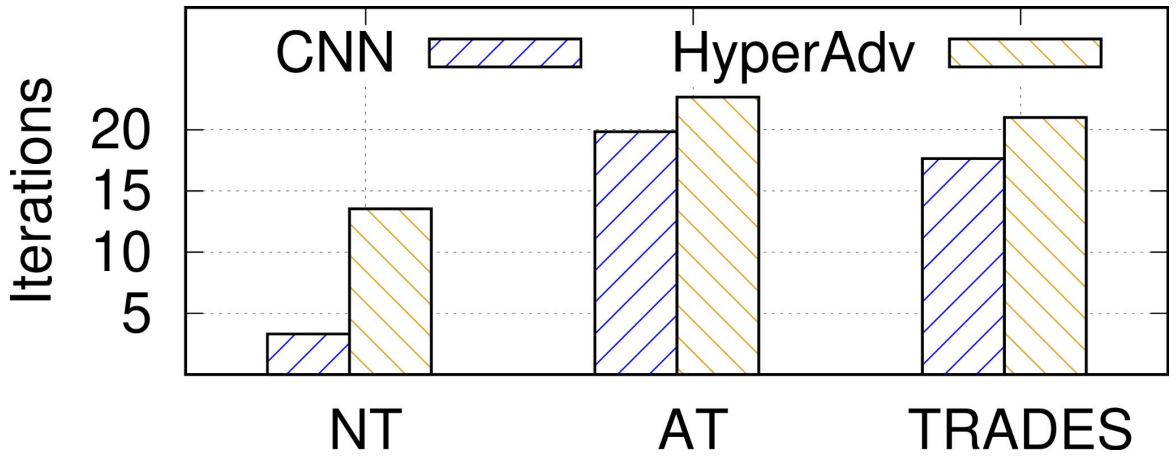
- The **dynamic** HyperAdv has best **robust** accuracy
- The **ensemble** HyperAdv performs best on **clean** data

Computational Cost of Adversary:

NT: 3.31 → 19.83 iterations

AT: 19.83 → 22.66 iterations

TRADES: 17.64 → 21.01 iterations



Future Work

Institute for the Wireless Internet of Things at Northeastern

- More scalable HyperNetwork design
- Latency and energy consumption
- Real-world implementation

Code and pretrained models are available:

https://github.com/Restuccia-Group/HyperAdv