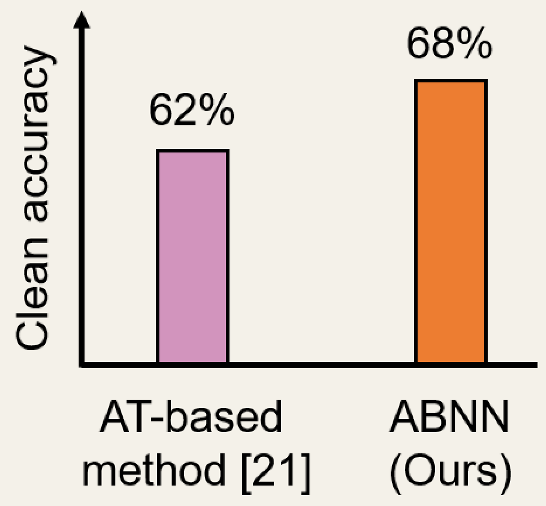
Adaptive Batch Normalization Networks for Adversarial Robustness

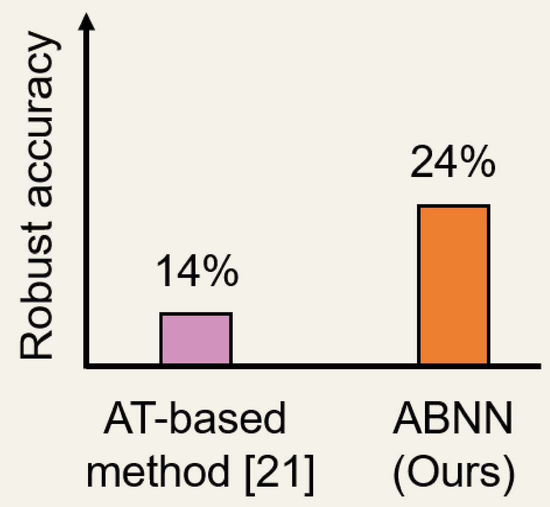
Shao-Yuan Lo and Vishal M. Patel
Johns Hopkins University

Contributions

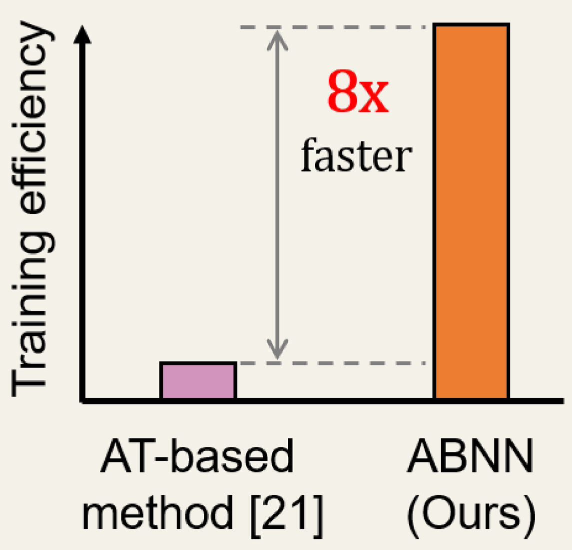
- We introduce a novel idea that uses test-time domain adaptation techniques to **defend against adversarial examples**.
- The proposed adversarial defense **ABNN** is a non-AT (Adversarial Training) method that **gets rid of the extremely time-consuming AT**.
- ABNN improves adversarial robustness against **both digital** and **physically realizable attacks** in **both image** and **video modalities**.
- Compared to traditional AT-based approaches, the proposed ABNN achieves **higher clean data performance**, **better robustness generalization**, and **significantly lower training time complexity**.



(a) Better clean data performance



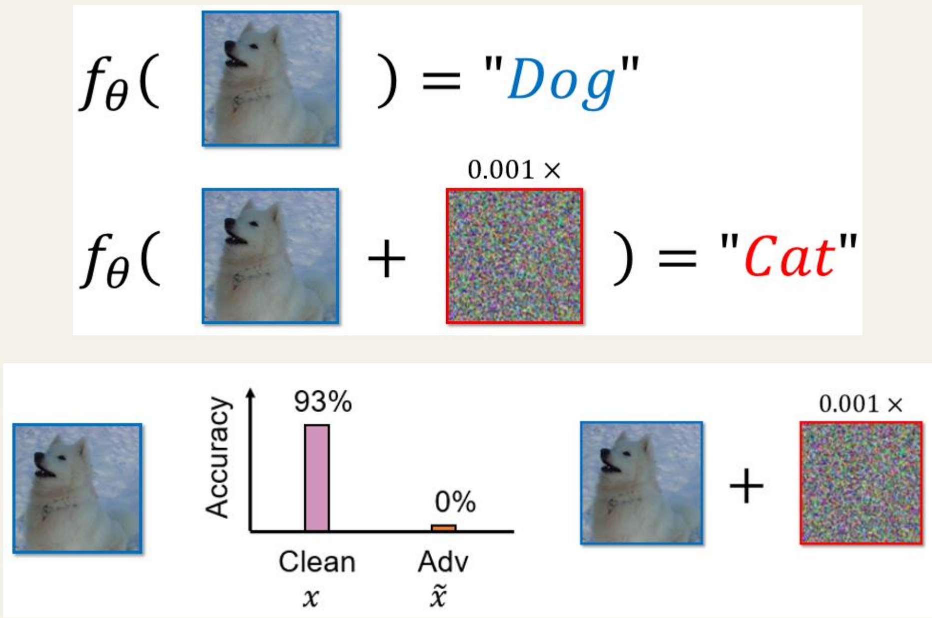
(b) Better robustness generalization



(c) Better training efficiency

Background

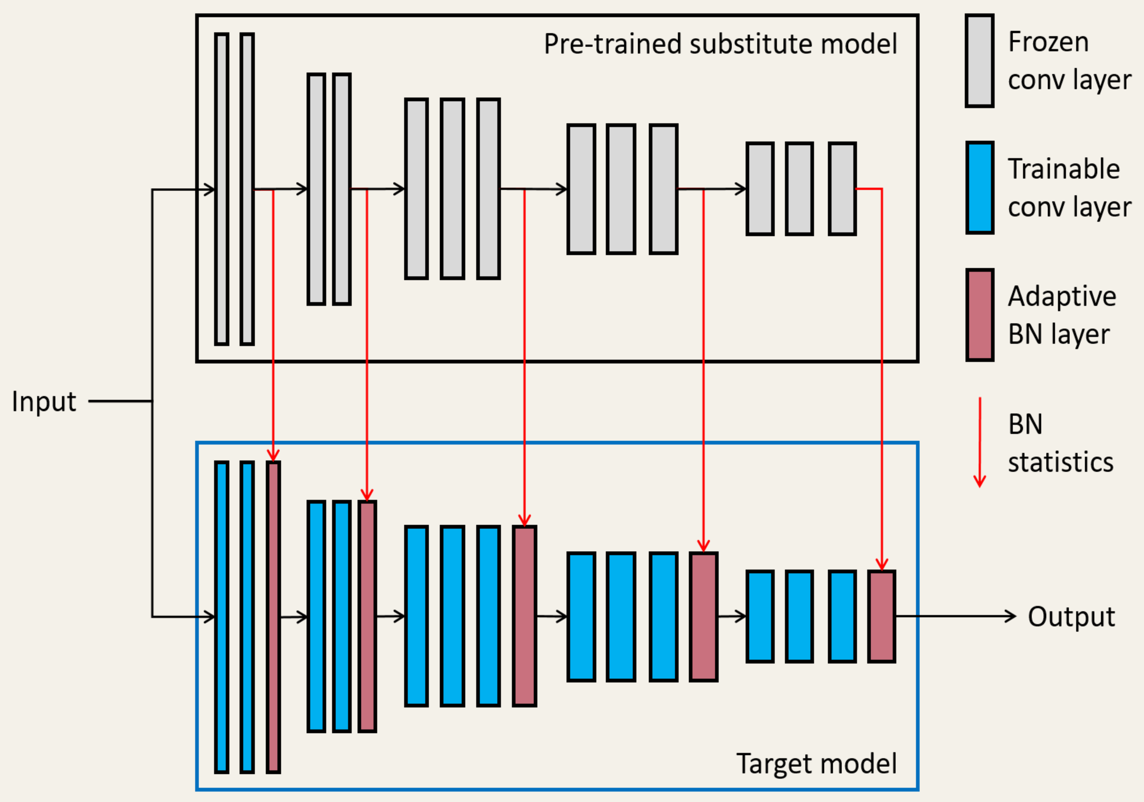
Deep networks are vulnerable to **adversarial examples**.



Adversarial Training (AT) is the most common defense method, but it involves min-max optimization, which is **extremely expensive**.

$$\underbrace{\theta^* = \arg\min_{\theta} \mathbb{E}_{(x,y)\sim\mathbb{D}} \Big[\underbrace{\max_{\delta\in\mathbb{S}} L(x+\delta, y; \theta)}_{\text{Generate adversarial examples with multiple iterations}} \Big]}_{\text{Train model parameters}}$$

Method



- **Pre-trained substitute model**: A public model trained on large-scale datasets (e.g., ImageNet)
- **Target model**: The model that we are training for a specific downstream task.
- **Standard Batch Normalization (BN)**:

$$z' = \gamma \left[\frac{z - \mu(z)}{\sigma(z)} \right] + \beta$$

- **The proposed adaptive BN**:

$$z'_t = \gamma_s \left[\sigma(z_s) \left[\frac{z_t - \mu(z_t)}{\sigma(z_t)} \right] + \mu(z_s) \right] + \beta_s$$

- **At training time**, the substitute model sends its corresponding BN statistics target model's adaptive BN layers, and the substitute model itself is frozen.
- **The model is trained on only clean data without AT**.
- **At test time**, our adaptive BN layer can adapt the substitute model's cleaner BN statistics to the target model, mitigating the adversarial effects in the target model's features.

Training Time Complexity

Let us set each network pass (i.e., a **forward** pass or a **backward** pass) to have **N** computational complexity, and let us suppose that ABNN's target network and substitute network have the same complexity.

- ABNN:
  2N + N = 3N
- PGT-AT:
  2N x tmax + 2N = 2N (tmax + 1)
- PGT-AT has
  2N (tmax + 1) / 3N = ***0.67 (tmax + 1)***
  times more training complexity than ABNN

Results

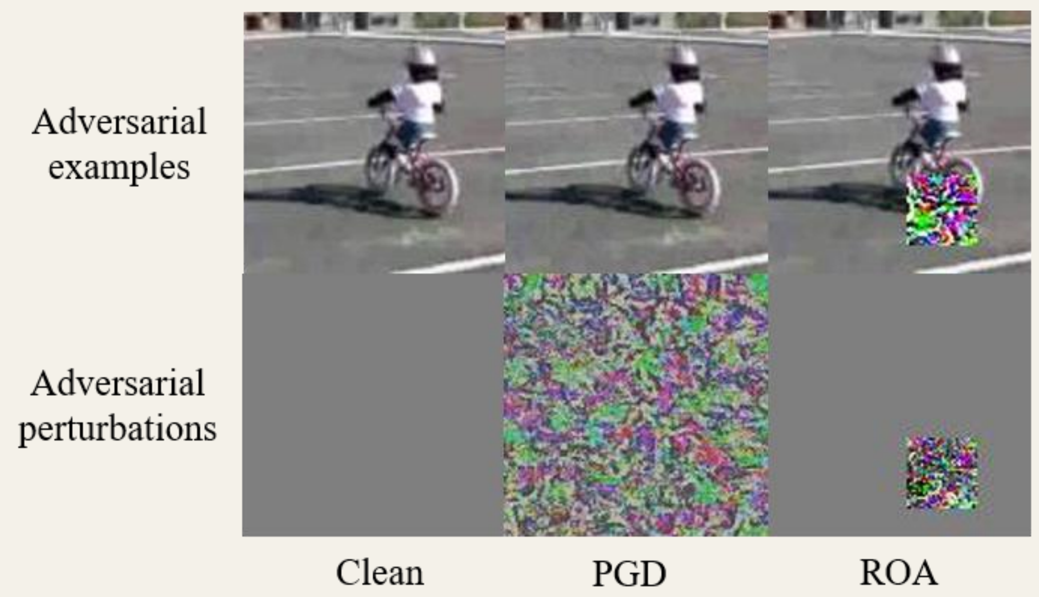


Table 1. Evaluation results (%) under the PGD attack on the CIFAR-10 dataset.

Method	Clean	PGD	Training cost
No Defense	93.4	0.0	2N
PGD-AT [23]	83.3	51.6	12N
ABNN (Ours)	87.5	31.5	3N

Table 2. Evaluation results (%) under the PGD attack on the UCF-101 dataset.

Method	Clean	PGD	Training cost
No Defense [14]	93.0	0.0	2N
OUDefend [21]	62.0	58.6	24N
ABNN (Ours)	68.3	43.4	3N

Table 3. Evaluation results (%) under the ROA attack on the UCF-101 dataset.

Method	Clean	ROA	Training cost
No Defense [14]	93.0	7.0	2N
OUDefend [21]	62.0	13.6	24N
ABNN (Ours)	68.3	24.4	3N

References

[14] K. A. Kinfu and R. Vidal, "Analysis and extensions of adversarial training for video classification," in CVPRW 2022.
[21] S.-Y. Lo, J. M. J. Valanarasu, and V. M. Patel, "Overcomplete representations against adversarial videos," in ICIP 2021.
[23] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in ICLR 2018.