Conserve-Update-Revise to Cure Generalization and Robustness Trade-off in Adversarial Training



Shruthi Gowda, Bahram Zonooz*, Elahe Arani*

TU/e EINDHOVEN UNIVERSITY OF TECHNOLOGY

ICLR

Introduction

Deep Neural Networks (DNNs) are ubiquitous in modern world, yet they are not without their limitations.

- Vulnerability to Adversarial Attacks - DNNs are susceptible to adversarial attacks, thus threatening the integrity and reliability of AI systems.

- Adversarial training is a promising strategy to enhance DNN robustness.

Challenges

- **Generalization and Robustness Trade-off:** Adversarial Training improves robustness but often compromises performance on clean images - Trade-off

- **Robust Overfitting:** Longer Adversarial Training can lead to reduced test performance.

Problem Statement

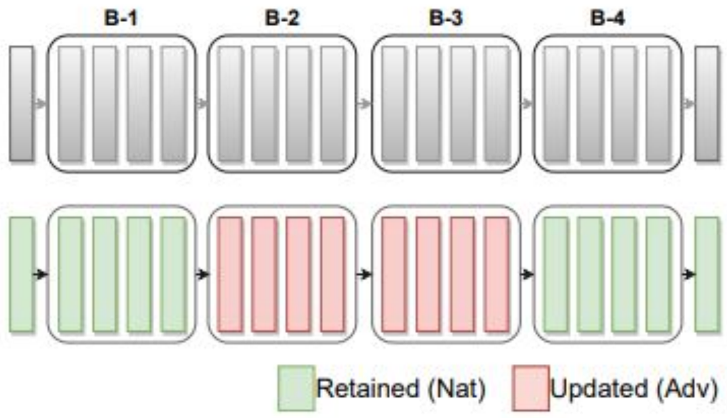
Understanding the Learning Dynamics: Exploring the learning patterns and capabilities of DNNs on both natural and adversarial data are crucial for reliable AI systems.

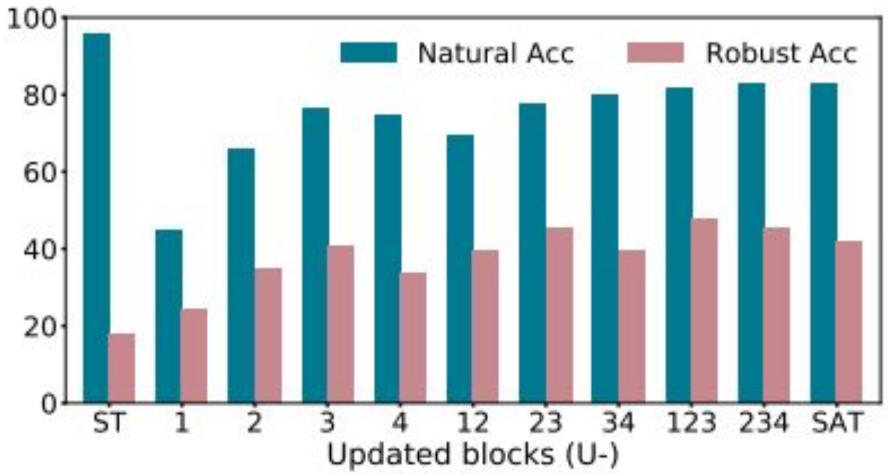
Perform an **Empirical Analysis** :
Investigate learning behavior during transition from Standard training to Adversarial training

- Layer-wise Analysis of weight updation and retention
- Representation similarity between features
- Overfitting phenomenon

Experimental Setup - Reinitialize different layers in each experiment while keeping the rest of the network fixed.

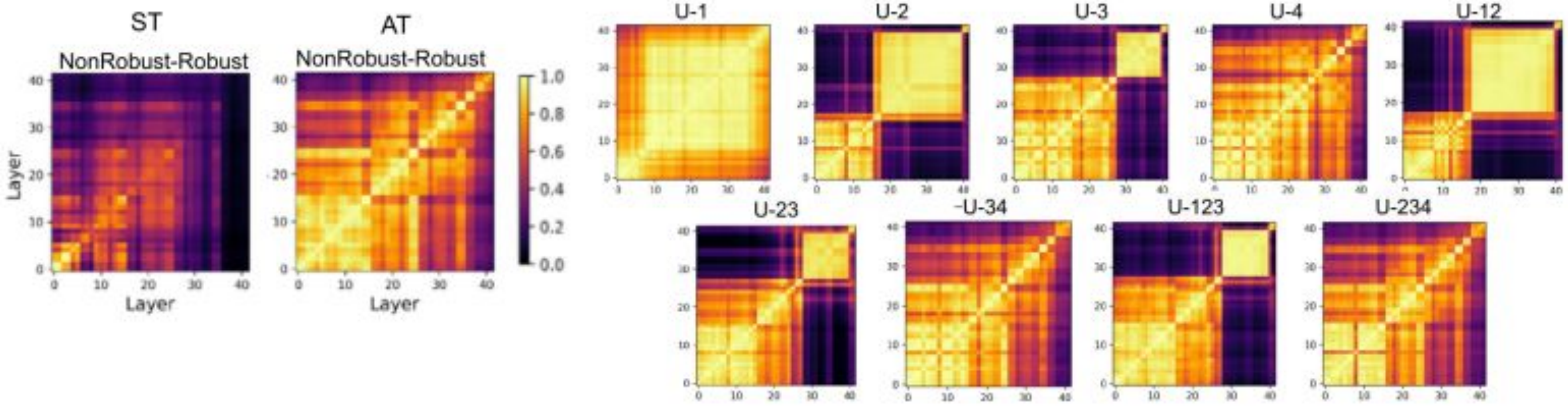
- The notation U-b represents the update of block ”b” while keeping the rest of the network frozen.



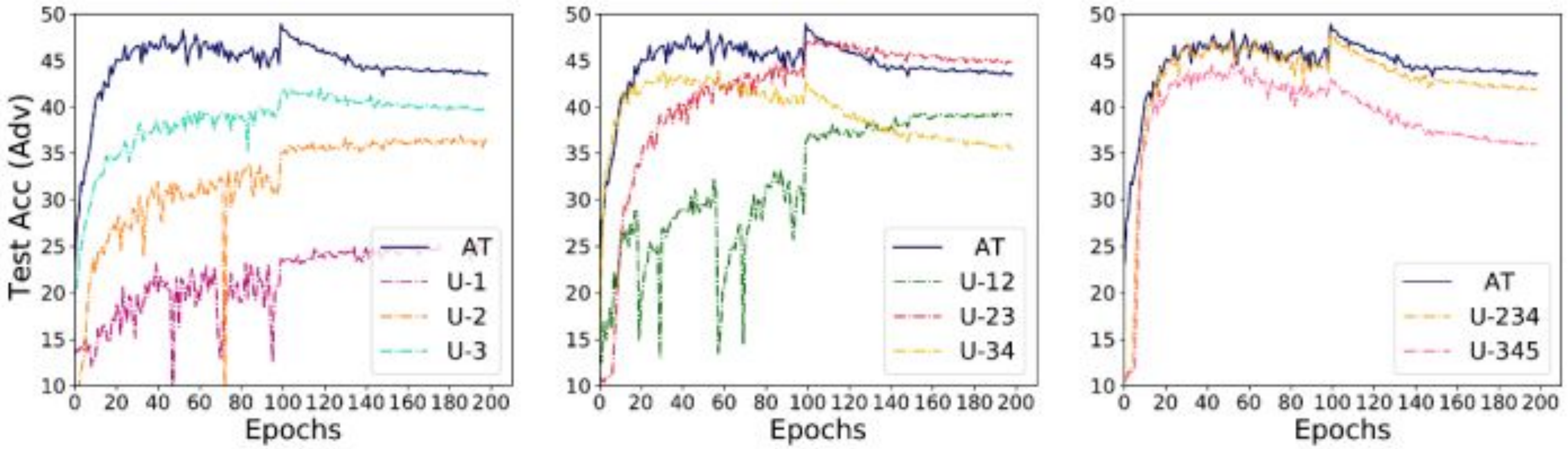


Standard generalization and robustness of different blocks of ResNet-18

Robust and Non-robust features - Visualizing features learned on natural and adversarial data aids in understanding representation alignment.



Over prolonged training in adversarial setting - test accuracy declines - **Robust Overfitting**
The base (AT) model prominently exhibits overfitting.



Methodology (1)

Empirical findings -

- Training the entire network and updating all weights may not be optimal for learning diverse data distributions.
- Selectively updating certain weights while conserving others can effectively leverage the network's learning capabilities.
- Retention and learning capabilities of the network - a better balance between natural and adversarial robustness

Propose a new Method : **CURE**

(1) **C**onservation (of knowledge from natural data),
(2) **U**pdation (of knowledge from adversarial data), and
(3) **RE**vision (of consolidated knowledge)

Methodology (2)

Adversarial Training
*

$$\delta^* = \arg\max_{\delta\in\Delta}\left[\mathcal{D}_{KL}(p(x_{nat};\theta)||p(x_{nat}+\delta;\theta))\right],$$

$$\mathcal{L}_{adv} = \mathcal{L}_{CE}(x_{nat};\theta) + \mathcal{D}_{KL}(p(x_{nat};\theta)||p(x_{adv},\theta)).$$

Robust Gradient Prominence (RGP) - determines which weights to update and which ones to freeze in

$$\mathcal{RGP}(w) = \alpha\left\|\frac{\partial\mathcal{L}(x_{nat};\theta)}{\partial w}\right\| + (1-\alpha)\left\|\frac{\partial\mathcal{L}(x_{adv};\theta)}{\partial w}\right\|$$

Revision stage - consolidate knowledge for Consistency regularization

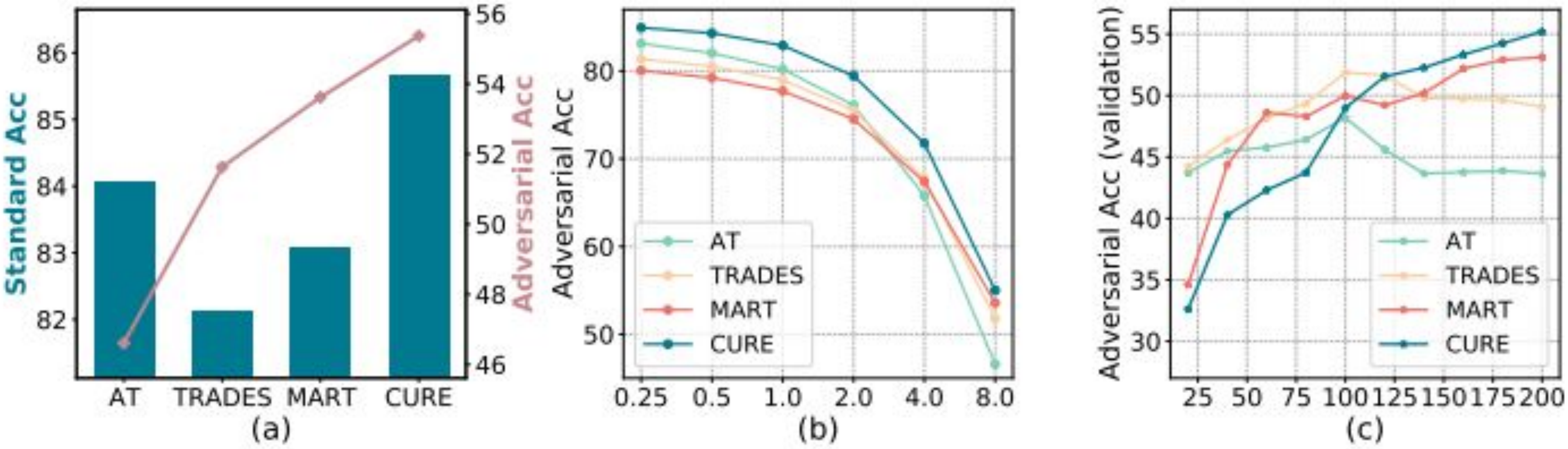
$$\mathcal{L}_{CR} = \mathcal{D}_{KL}(p(x_{nat};\theta_{rev})||p(x_{nat};\theta)) + \mathcal{D}_{KL}(p(x_{adv};\theta_{rev})||p(x_{adv};\theta)).$$

Results (1)

Method		WideResNet-34-10					ResNet-18				
		Nat	PGD20	AA	C&W	NRR	Nat	PGD20	AA	C&W	NRR
AT	ICLR'18	85.17	55.08	44.04	52.91	65.27	82.78	51.30	44.63	49.72	62.12
TRADES	ICML'19	84.73	56.82	52.95	54.29	66.17	82.41	52.76	48.37	50.43	62.57
MART	ICLR'20	83.62	56.74	51.23	53.16	64.99	80.70	54.02	47.49	49.35	61.24
FAT	ICML'20	86.60	49.86	47.48	49.35	62.87	87.72	46.69	43.14	49.66	63.41
ST-AT	ICLR'23	84.92	57.73	53.54	-	-	83.10	54.62	50.50	51.43	63.53
ACT	BMVC'20	87.10	54.77	-	-	-	84.33	55.83	-	-	-
ARD	AAAI'20	85.18	53.79	-	-	-	82.84	51.41	-	-	-
IAD	ICLR'22	83.06	56.17	52.68	53.99	65.44	80.63	53.84	50.17	51.60	62.92
LAS-AT	CVPR'22	85.24	57.07	53.58	55.45	67.19	82.39	53.70	49.94	51.96	63.72
CURE	-	87.05	58.28	52.10	55.25	**67.60**	86.76	54.92	49.69	52.48	**65.04**

$$NRR = \frac{2 \times Natural Accuracy \times Robust Accuracy}{Natural Accuracy + Robust Accuracy}$$

Results (2)



Generalization-Robustness TradeOff

Performance against different perturbation strengths

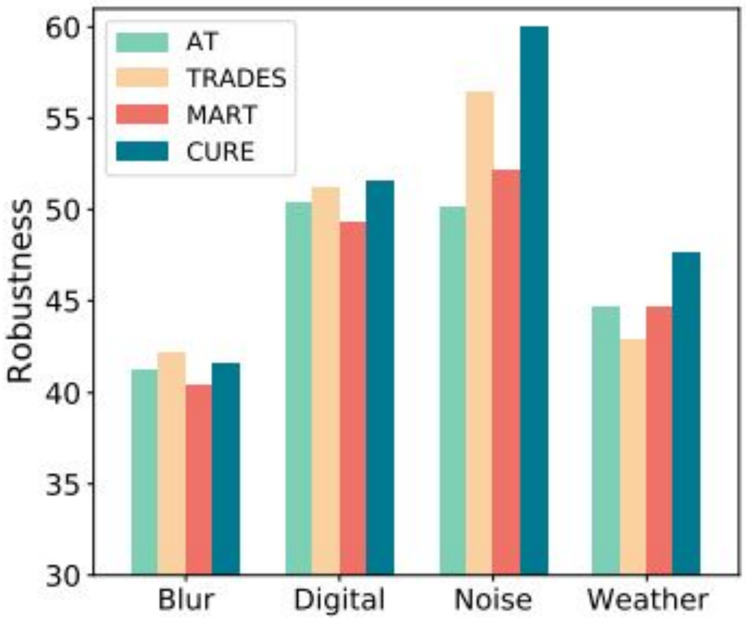
Robust Overfitting

Results (3)

- DNNs are vulnerable to natural corruptions

- Figure illustrates CURE's effectiveness in addressing multiple types of corruptions

- CURE showcases improved resistance and stability compared to traditional methods.
    -

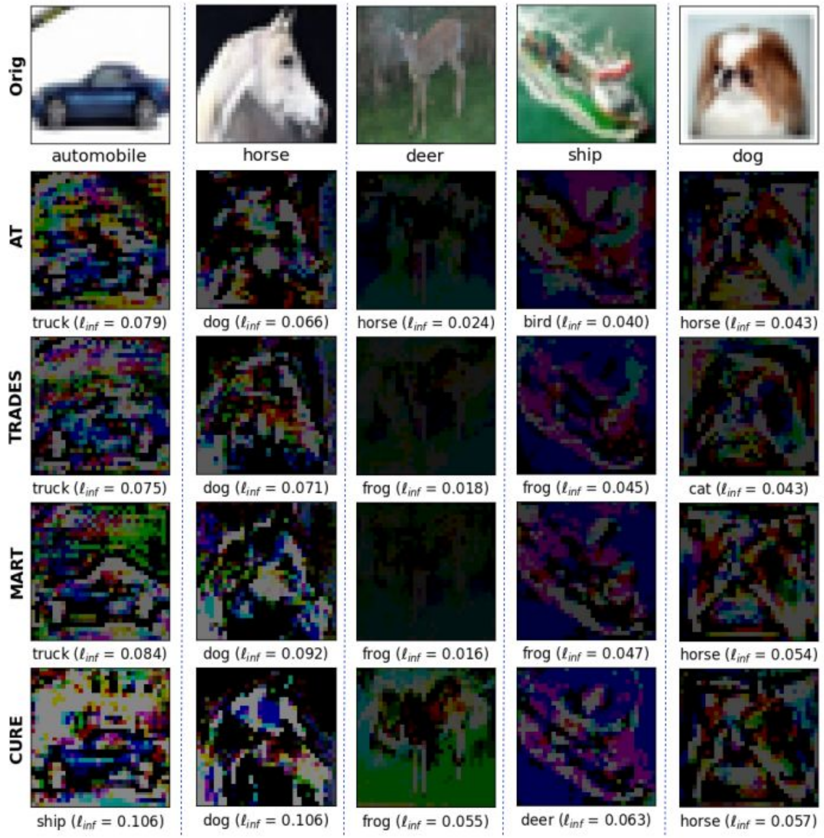


Results (4)

Adversarial Perturbations

- The visualizations provide a clear comparison of the minimum perturbations required to fool each of the robust models

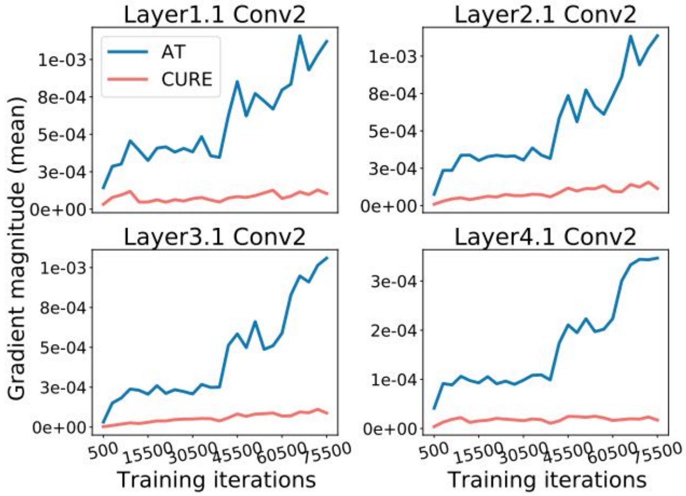
- Models trained with CURE exhibit a higher level of sensitivity to perturbations.

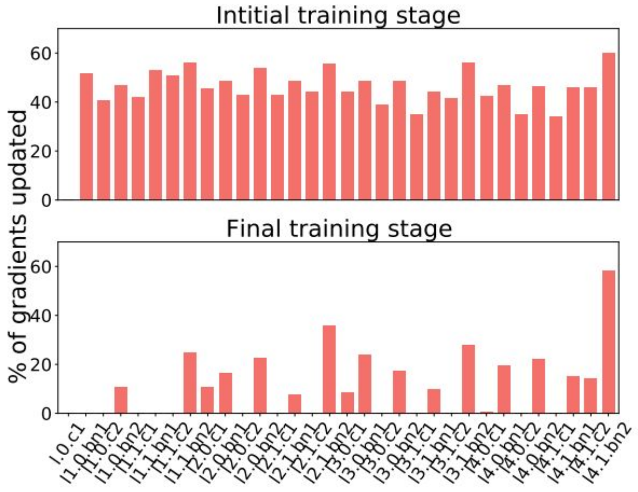


Results (5)

Gradients Analysis

- Percentage of gradients updated in each layer conv layer during initial and final phases of training
- As training progresses, the RGP metric identifies the weights that need to be fixed to prevent overwriting.





THANKS

Contact: Shruthi Gowda
Email: s.gowda@tue.nl
Website: https://github.com/NeurAI-Lab