Towards Efficient and Effective Adversarial Training




Gaurang Sriramanan*   Sravanti Addepalli*   Arya Baburaj   R. Venkatesh Babu

Video Analytics Lab, Department of Computational and Data Sciences
Indian Institute of Science, Bangalore

Introduction

NuAT: Nuclear Norm Adversarial Training

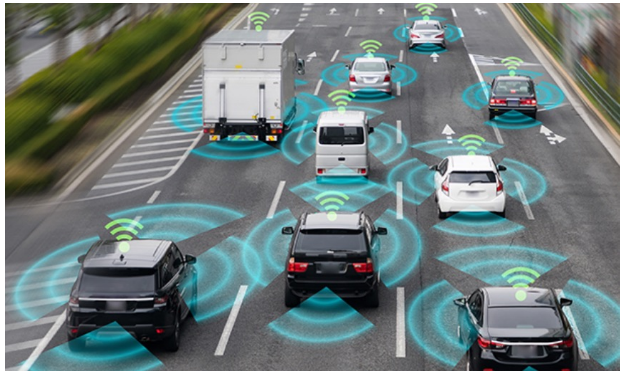
Experiments and Analysis

Summary

Introduction

Deep Learning Applications

- Autonomous navigation systems
- Surveillance systems
- Medicine and health care
- Reinforcement learning
- Generative modelling
- Style transfer
- Robotics
- Speech Processing
- Natural Language Processing



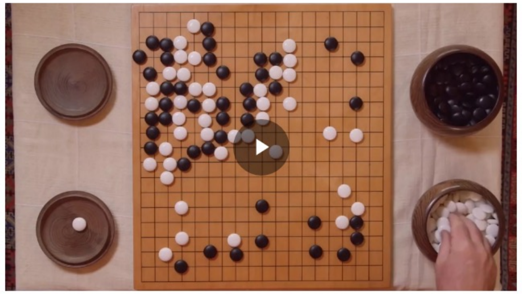
https://www.theparliamentmagazine.eu/news/article/autonomous-driving-a-glimpse-into-the-future

Google's DeepMind defeats legendary Go player Lee Se-dol in historic victory

39

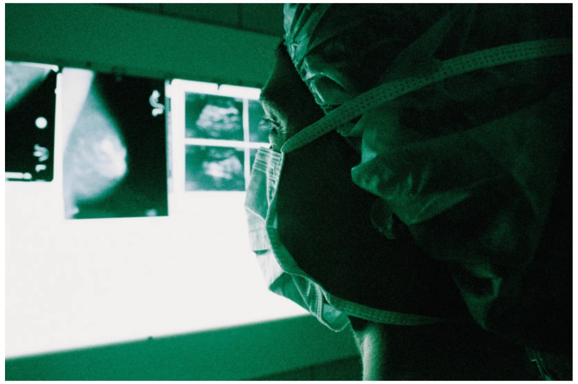
By Sam Byford | @345triangle | Mar 9, 2016, 2:32am EST

SHARE

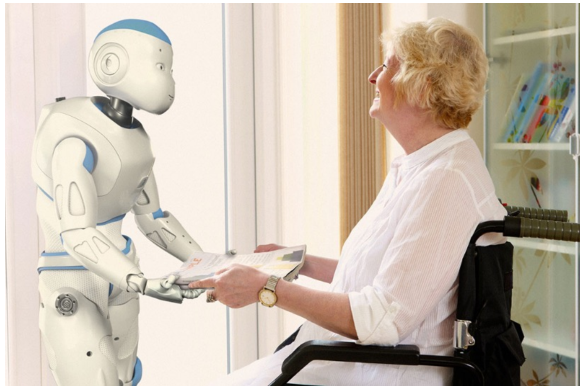


https://www.theverge.com/2016/3/9/11184362/google-alphago-go-deepmind-result

AI beats docs in cancer spotting



https://paulbiegler.com/2017/12/21/ai-beats-docs-in-cancer-spotting/



https://www.icsfoundation.ie/can-make-care-robots-affordable-need/

Adversarial Attacks

Prediction: **Hamster**

Confidence = 99.99%

$+ 0.02 \ *$

50-step PGD targeted attack with $\varepsilon = \frac{8}{255}$ scaled by 50x

$=$

Prediction: **Banjo**

Confidence = 100%

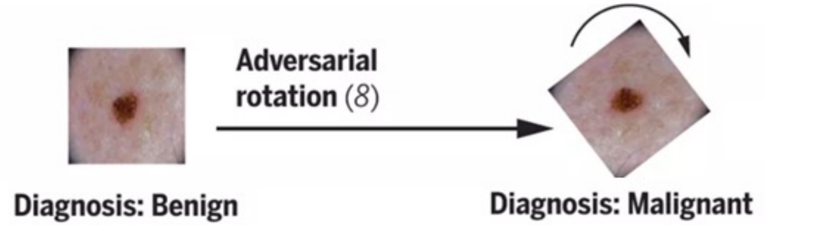
Motivation for Adversarial Defense Research

Hackers can trick a Tesla into accelerating by 50 miles per hour

A two inch piece of tape fooled the Tesla's cameras and made the car quickly and mistakenly speed up.

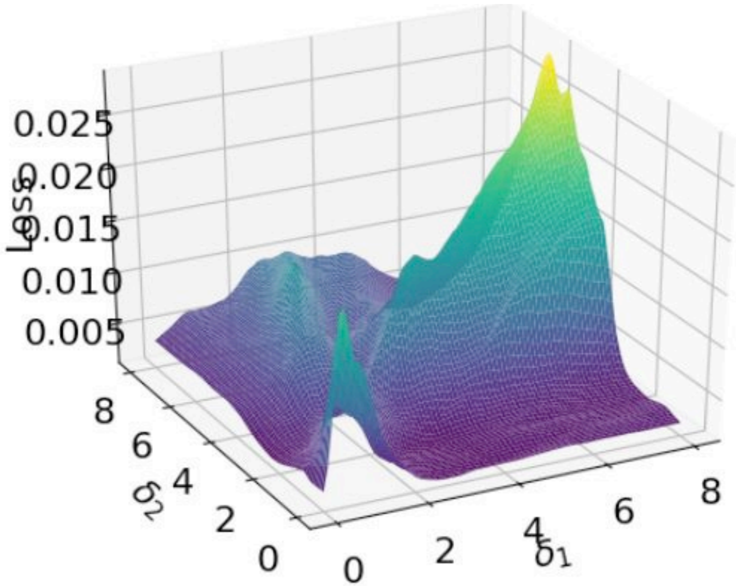
https://www.technologyreview.com/2020/02/19/868188/hackers-can-trick-a-tesla-into-accelerating-by-50-miles-per-hour/

Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition, M Sharif, S Bhagavatula, L Bauer, MK Reiter, ACM SIGSAC 2016



https://www.vox.com/future-perfect/2019/4/8/18297410/ai-tesla-self-driving-cars-adversarial-machine-learning

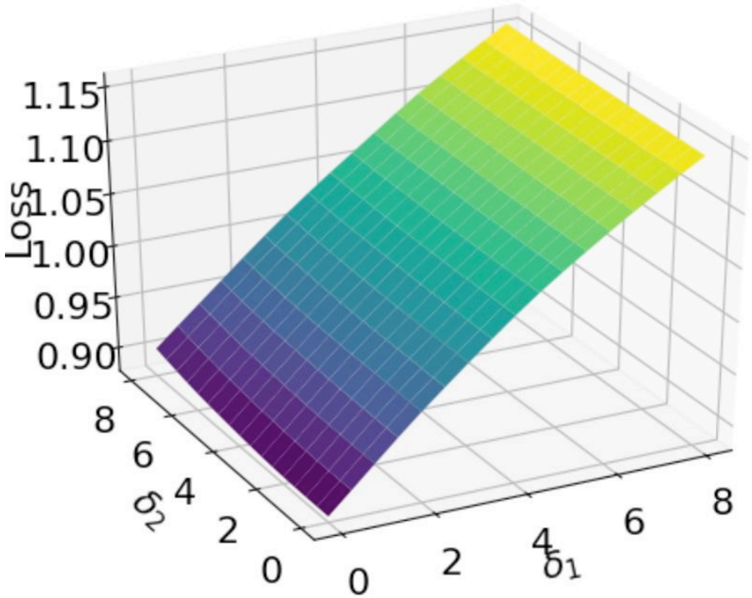
Defending against Adversarial Attacks



FGSM-AT

Single-step defenses

- Single-step gradients used for attack generation
- FGSM training [2]
- Low computational cost
- Susceptible to Gradient Masking leading to a false sense of security and training instability
- Suboptimal clean accuracy and robustness



NuAT (Ours)

[1] Guo et al. Countering adversarial images using input transformations. ICLR, 2018.
[2] Goodfellow et al. Explaining and harnessing adversarial examples. ICLR, 2015.
[3] Madry et al. Towards Deep Learning Models Resistant to Adversarial Attacks. ICLR, 2018.
[4] Zhang et al. Theoretically principled trade-off between robustness and accuracy. ICML, 2019.

NuAT: Nuclear Norm Adversarial Training

Preliminaries: Nuclear Norm

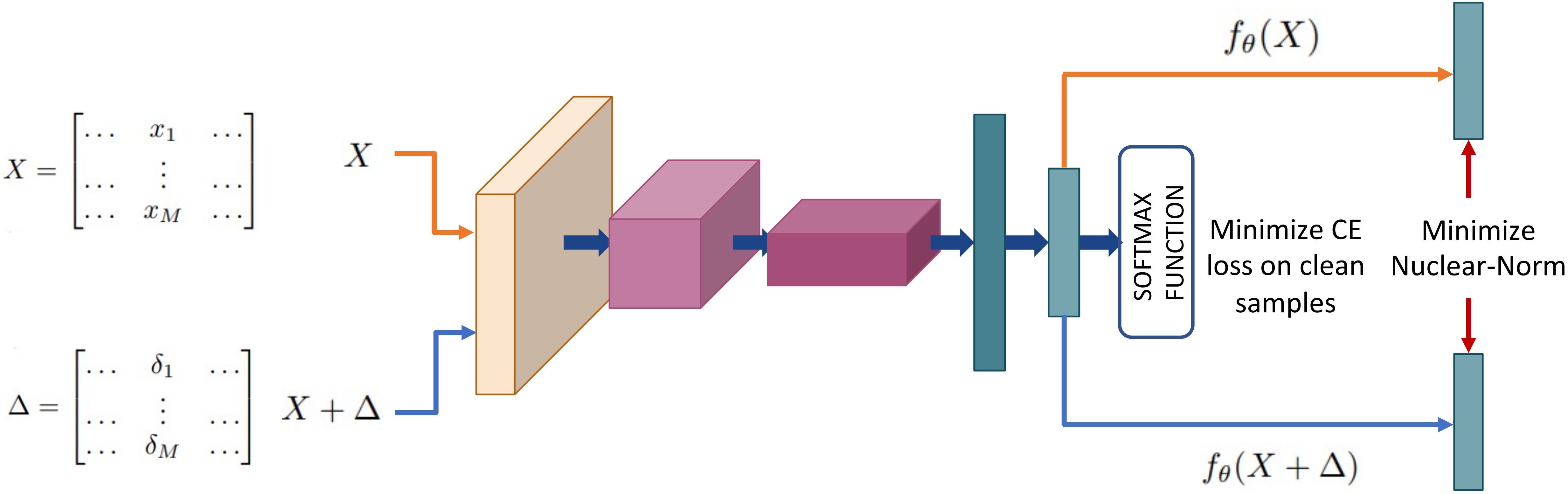
$$\|A\|_* = \sum_{i=1}^{\min\{m,n\}} \sigma_i(A) = \mathrm{trace}\left(\sqrt{A^*A}\right)$$

- Forms a uniform upper bound of the Frobenius Norm
- Let $A = U\Lambda V^T$ be the Singular Value Decomposition of A

$$||A||_*^2 = \left(\sum_{i=1}^{\rho} \sigma_i\right)^2 = \sum_{i=1}^{\rho} \sigma_i^2 + \sum_{i \neq j} \sigma_i \cdot \sigma_j \geq \sum_{i=1}^{\rho} \sigma_i^2$$

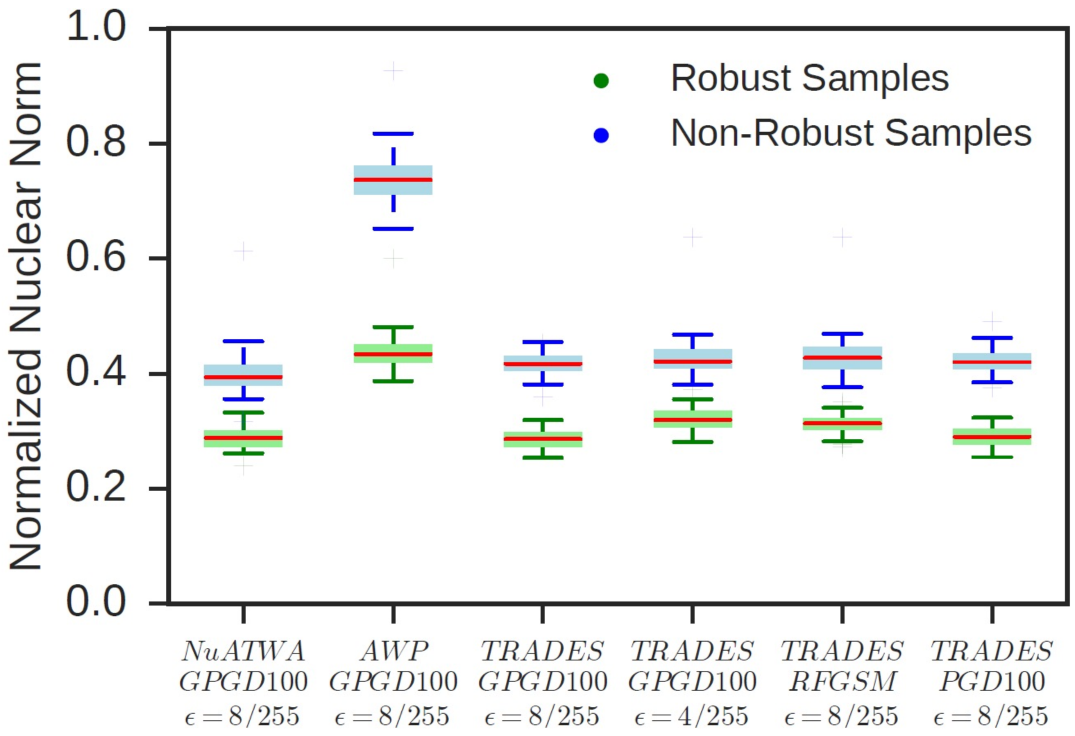
$$\sum_{i=1}^{\rho} \sigma_i^2 = ||\Lambda||_F^2 = ||U\Lambda||_F^2 = ||U\Lambda V^T||_F^2 = ||A||_F^2$$

Nuclear Norm Regularization



$$L = \ell_{CE}(f_\theta(X), Y) + \lambda \cdot ||f_\theta(\widetilde{X}) - f_\theta(X)||_*$$

Nuclear Norm Regularization



Generation of Nuclear-Norm based attack

For a training minibatch $B = \{(x_i, y_i)\}_{i=1}^{M}$,

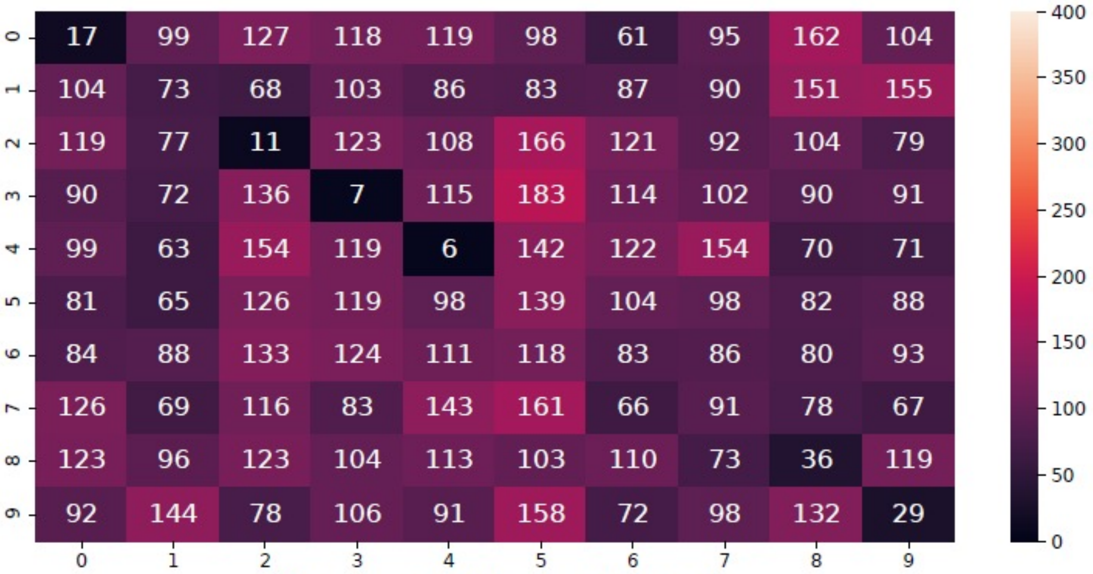
$$X = \begin{bmatrix} \cdots & x_1 & \cdots \\ \cdots & \vdots & \cdots \\ \cdots & x_M & \cdots \end{bmatrix}, \; Y = \begin{bmatrix} y_1 \\ \vdots \\ y_M \end{bmatrix} \quad \Delta = \begin{bmatrix} \cdots & \delta_1 & \cdots \\ \cdots & \vdots & \cdots \\ \cdots & \delta_M & \cdots \end{bmatrix}, \quad \delta_i \sim Bern^d(-\alpha, \alpha)$$

$$\widetilde{L} = \ell_{CE}\left(f_\theta(X + \Delta), Y\right) + \lambda \cdot ||f_\theta(X + \Delta) - f_\theta(X)||_*$$

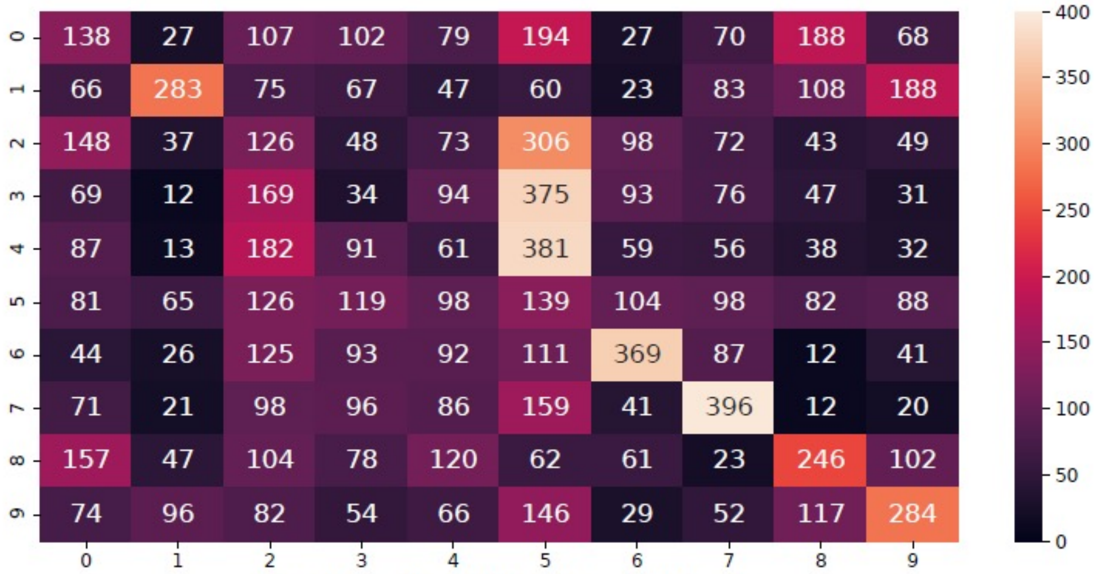
$$\Delta = \Delta + \varepsilon \cdot \text{sign}\left(\nabla_\Delta \widetilde{L}\right)$$

$$\Delta = Clamp\left(\Delta, -\varepsilon, \varepsilon\right), \quad \widetilde{X} = Clamp\left(X + \Delta, 0, 1\right)$$

Diversity of Nuclear-Norm Attack



Nuclear Norm Attack



Frobenius Norm Attack

Confusion Matrices for predictions against adversarial attacks generated by maximizing the Nuclear norm and Frobenius norm of a matrix respectively. These are obtained for a normally trained model with ResNet-18 architecture on CIFAR-10 dataset.

NuAT: Nuclear-Norm Adversarial Training

Repeat for I iterations

Single-step Nuclear Norm based attack

$$\widetilde{L} = \ell_{CE}\left(f_\theta(X + \Delta), Y\right) + \lambda \cdot ||f_\theta(X + \Delta) - f_\theta(X)||_*$$

$$\Delta = \Delta + \varepsilon \cdot \mathrm{sign}\left(\nabla_\Delta \widetilde{L}\right)$$

$$\Delta = Clamp\left(\Delta, -\varepsilon, \varepsilon\right), \quad \widetilde{X} = Clamp\left(X + \Delta, 0, 1\right)$$

Adversarial Training

$$L = \ell_{CE}(f_\theta(X), Y) + \lambda \cdot ||f_\theta(\widetilde{X}) - f_\theta(X)||_*$$

Parameter update

$$\theta = \theta - \frac{1}{M} \cdot \eta \cdot \nabla_\theta L$$

NuAT-WA

Repeat for I iterations

Single-step Nuclear Norm based attack

$$\widetilde{L} = \ell_{CE}\left(f_\theta(X+\Delta), Y\right) + \lambda \cdot ||f_\theta(X+\Delta) - f_\theta(X)||_*$$

$$\Delta = \Delta + \varepsilon \cdot \mathrm{sign}\left(\nabla_\Delta \widetilde{L}\right)$$

$$\Delta = Clamp\left(\Delta, -\varepsilon, \varepsilon\right), \quad \widetilde{X} = Clamp\left(X+\Delta, 0, 1\right)$$

Adversarial Training

$$L = \ell_{CE}(f_\theta(X), Y) + \lambda \cdot ||f_\theta(\widetilde{X}) - f_\theta(X)||_*$$

Parameter update

$$\theta = \theta - \frac{1}{M} \cdot \eta \cdot \nabla_\theta L, \quad \omega = (1-\tau) * \theta + \tau * \omega$$

NuAT2: 2-step Adversarial Training

First attack step

$$\widetilde{L} = \ell_{CE}\left(f_\theta(X+\Delta), Y\right) + \lambda \cdot ||f_\theta(X+\Delta) - f_\theta(X)||_*$$

$$\Delta = \Delta + \varepsilon \cdot \text{sign}\left(\nabla_\Delta \widetilde{L}\right)$$

$$\Delta = Clamp\left(\Delta, -\varepsilon, \varepsilon\right), \quad \widetilde{X} = Clamp\left(X+\Delta, 0, 1\right)$$

Second attack step

$$\Delta = \Delta + \varepsilon \cdot \text{sign}(\nabla_\Delta \ell_{CE}(f_\theta(X+\Delta), Y))$$

$$\Delta = Clamp\left(\Delta, -\varepsilon, \varepsilon\right), \quad \widetilde{X} = Clamp\left(X+\Delta, 0, 1\right)$$

NuAT2-WA

First attack step from EMA (Exponential moving average) model

$$\widetilde{L} = \ell_{CE}(f_\omega(X+\Delta), Y) + \lambda \cdot ||f_\omega(X+\Delta) - f_\omega(X)||_*$$

$$\Delta = \Delta + \varepsilon \cdot \text{sign}\left(\nabla_\Delta \widetilde{L}\right)$$

$$\Delta = Clamp\,(\Delta, -\varepsilon, \varepsilon)$$

Second attack step from the model being trained

$$\Delta = \Delta + \varepsilon \cdot \text{sign}(\nabla_\Delta \ell_{CE}(f_\theta(X+\Delta), Y))$$

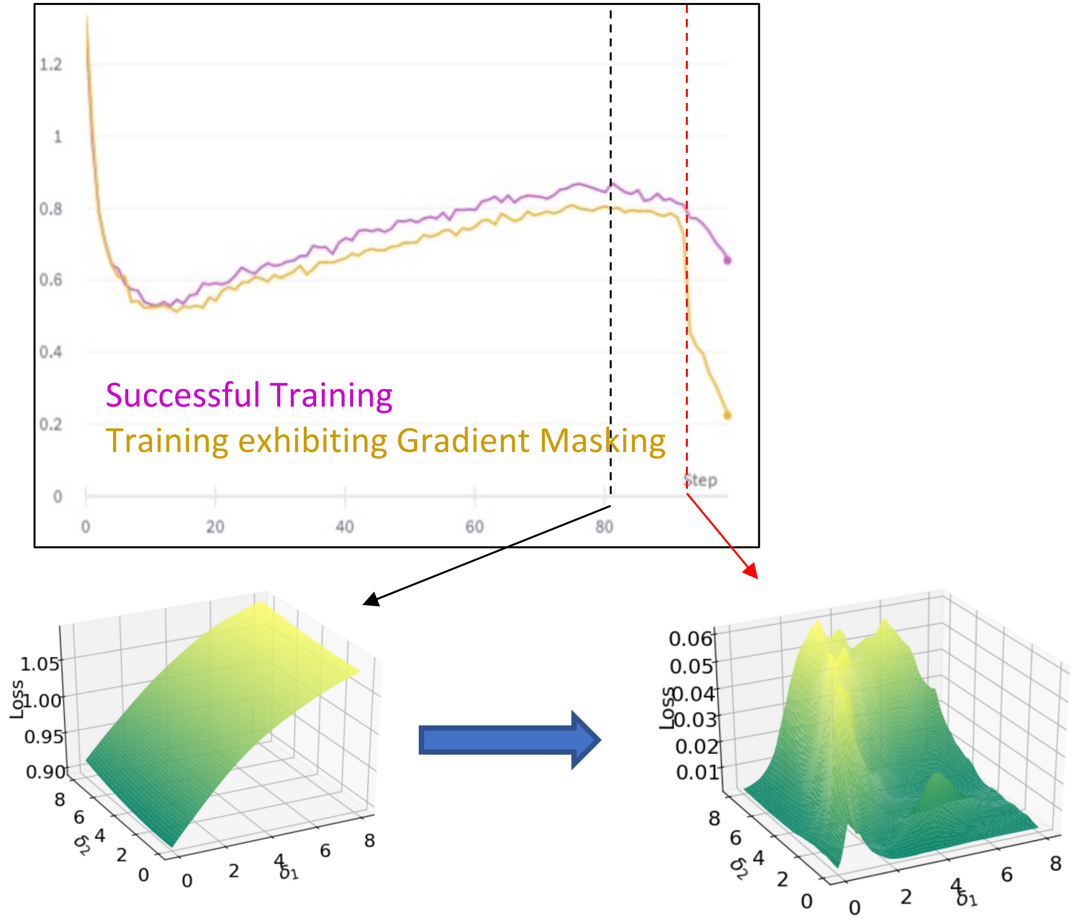
$$\Delta = Clamp\,(\Delta, -\varepsilon, \varepsilon), \quad \widetilde{X} = Clamp\,(X+\Delta, 0, 1)$$

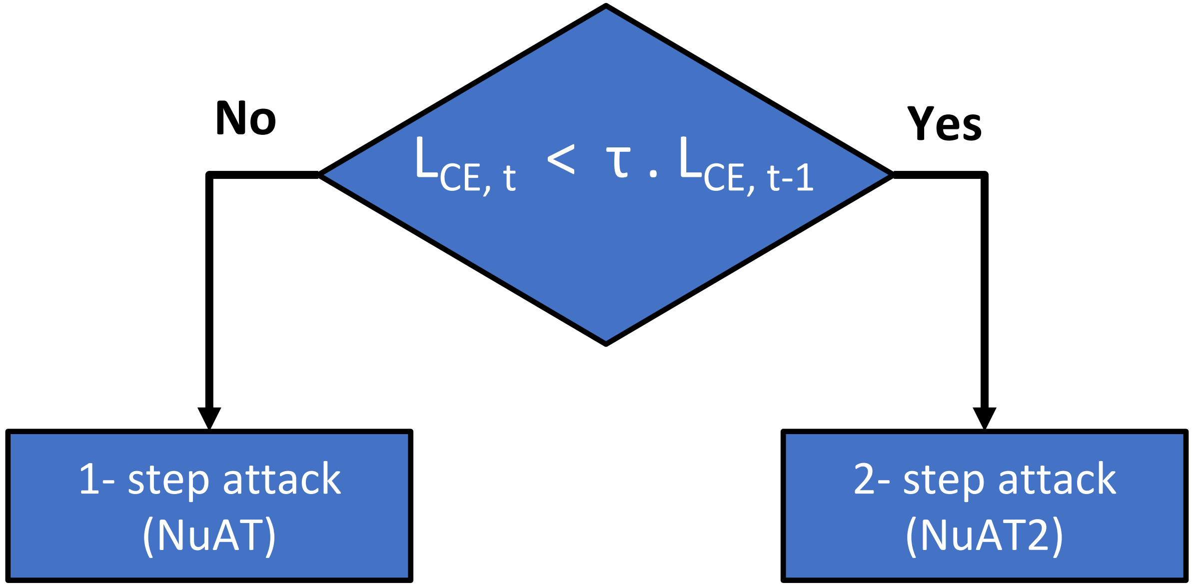
Update weights of EMA model

$$\omega = (1-\tau) * \theta + \tau * \omega$$

Hybrid Adversarial Training (NuAT-H)

Cross-Entropy Loss (Clean)





B. Li, S. Wang, S. Jana, and L. Carin. Towards understanding fast adversarial training. arXiv preprint, arXiv:2006.03089, 2020

Experiments and Analysis

Results on CIFAR-10 (ResNet-18)

Method	# AT steps	Clean Acc	PGD (n-steps) 20	PGD (n-steps) 500	GAMA 100	AA (v1)
Normal	0	92.30	0.00	0.00	0.00	0.00
FGSM-AT	1	**92.89**	0.00	0.00	0.00	0.00
RFGSM-AT	1	89.24	35.02	34.17	33.87	33.16
ATF	1	71.77	43.53	43.52	40.34	40.22
FBF	1	82.83	46.41	46.03	43.85	43.12
R-MGM	1	82.29	46.23	45.79	44.06	43.72
GAT	1	80.49	53.13	53.08	47.76	47.30
GAT-WA	1	79.47	**54.40**	**54.37**	49.00	48.28
NuAT **(Ours)**	1	81.01	53.30	52.97	49.46	49.24
NuAT-WA **(Ours)**	1	82.21	54.14	53.95	**50.97**	**50.75**
PGD-AT	10	81.12	53.08	52.89	49.08	48.75
TRADES	10	81.47	52.73	52.61	49.22	49.06
TRADES-WA	10	80.19	52.98	52.88	49.49	49.39
AWP	11	**81.99**	**55.60**	**55.52**	**51.65**	**51.45**

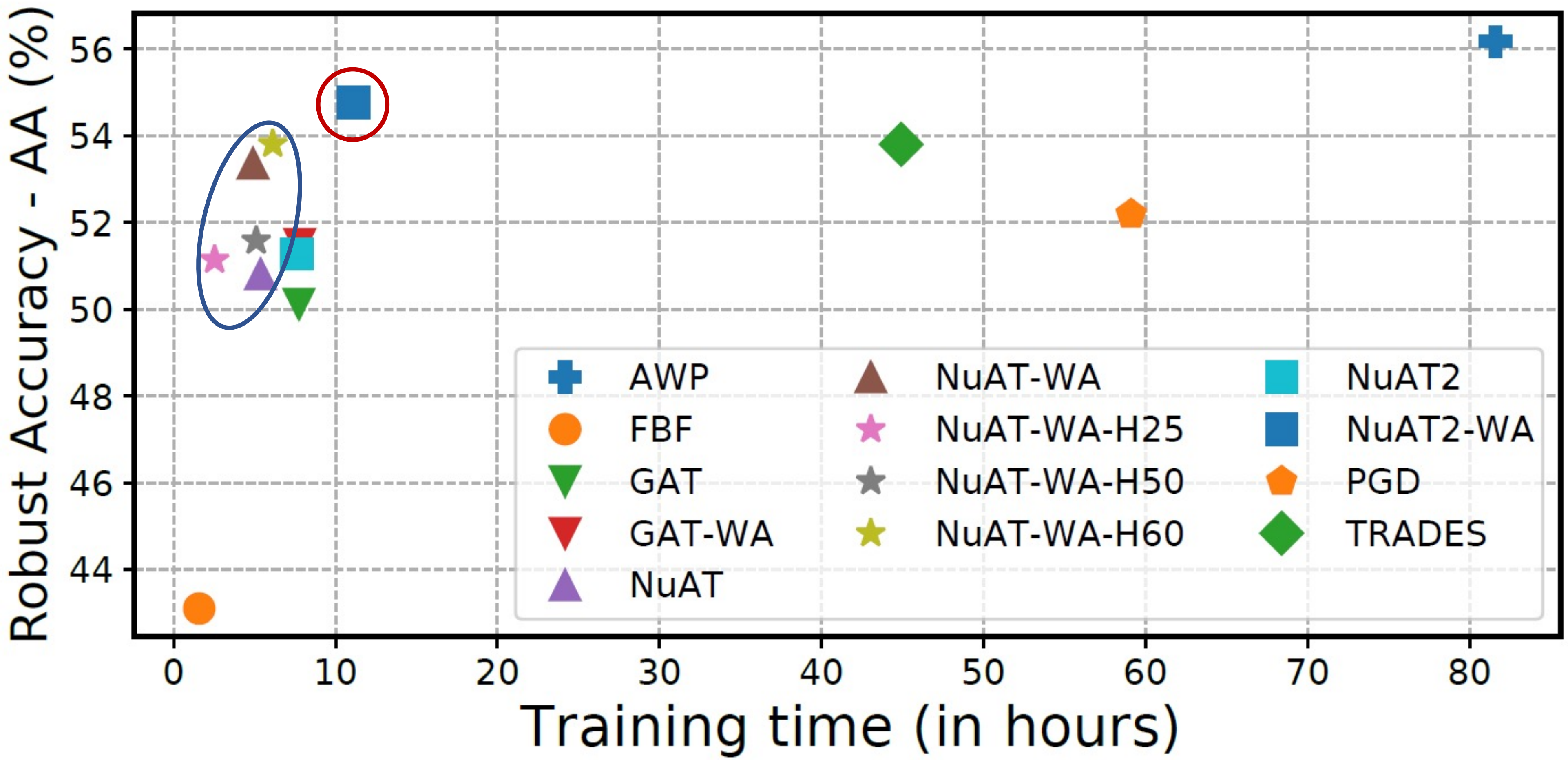
Results on CIFAR-10 (WideResNet-34-10)

Method	**AT-steps (epochs)**	**Clean Acc**	**PGD 100**	**GAMA 100**	**AA (v2)**
FBF	1 (30)	82.05	45.57	43.13	43.10
GAT	1 (85)	85.17	55.12	50.76	50.12
GAT-WA	1 (85)	84.61	57.28	52.19	51.50
Variants of NuAT (**Ours**)					
NuAT	1 (55)	85.30	53.82	51.34	50.81
NuAT-H	1 (50_{+5})	84.58	54.89	51.93	51.58
NuAT-WA	1 (50)	85.29	56.21	53.73	53.36
NuAT-WA-H	1 (25_{+2})	81.98	54.82	51.41	51.14
NuAT-WA-H	1 (60_{+6})	84.93	57.51	54.28	53.81
NuAT2	2 (55)	84.76	54.50	51.99	51.27
NuAT2-WA	2 (80)	**86.32**	**57.74**	**55.08**	54.76
TRADES	10 (110)	85.48	56.35	53.88	53.80
PGD	10 (200)	**86.07**	55.74	52.70	52.19
AWP	11 (200)	85.36	**59.13**	**56.35**	**56.17**

Results across different datasets

	CIFAR-10				ImageNet-100				MNIST			
	Clean Acc	**PGD 500**	**GAMA 100**	**AA (v1)**	**Clean Acc**	**PGD 500**	**GAMA 100**	**AA (v1)**	**Clean Acc**	**PGD 500**	**GAMA 100**	**AA (v1)**
Normal	**92.30**	0.00	0.00	0.00	**81.44**	0.00	0.00	0.00	99.20	0.00	0.00	0.00
RFGSM-AT	89.24	34.17	33.87	33.16	78.46	13.88	13.38	12.96	**99.37**	85.32	83.64	82.28
FBF	82.83	46.03	43.85	42.37	57.32	27.22	21.78	20.66	99.30	91.37	87.27	79.02
R-MGM	82.29	45.79	44.06	43.72	64.84	31.68	27.46	27.68	99.04	90.56	88.13	86.21
GAT	80.49	53.08	47.76	47.30	67.98	37.46	29.30	28.92	**99.37**	94.44	92.96	90.62
NuAT (**Ours**)	81.01	52.97	49.46	49.24	69.00	37.60	32.38	31.96	**99.37**	96.24	94.65	**93.11**
NuAT-WA (**Ours**)	82.21	**53.95**	**50.97**	**50.75**	68.40	**38.68**	**33.22**	**33.16**	99.36	**96.30**	**94.70**	93.10
TRADES	81.47	52.61	49.22	49.06	62.88	37.24	31.44	31.66	99.32	93.40	92.74	92.19
PGD-AT	81.12	52.89	49.08	48.75	68.62	36.56	32.24	32.98	99.27	93.98	92.80	91.81

Efficiency and Effectiveness of NuAT



Summary

Nuclear Norm Attack

Frobenius Norm Attack

Robust Accuracy - AA (%)
Training time (in hours)
AWP
FBF
GAT
GAT-WA
NuAT
NuAT-WA
NuAT-WA-H25
NuAT-WA-H50
NuAT-WA-H60
NuAT2
NuAT2-WA
PGD
TRADES

Loss

Summary

- Nuclear Norm Adversarial Training (NuAT) to improve adversarial robustness at low computational cost
- **NuAT**: SOTA across various single-step defenses
- **NuAT2**: Achieves results better than some multi-step (10-step) defenses (TRADES, PGD-AT), and comparable to the SOTA defense, TRADES-AWP
- **NuAT-H**: Bridges the computation-accuracy trade-off between NuAT and NuAT2
- Scales to large network capacities such as WideResNet
- Scales to large datasets such as ImageNet-100.

Thank You!

Acknowledgements: This work was supported by Uchhatar Avishkar Yojana (UAY) project (IISC 10), MHRD, Govt. of India. Sravanti Addepalli is supported by a Google PhD Fellowship in Machine Learning. We are thankful for the support.