Adversarial Robustness against Multiple and Single l_p-Threat Models via Quick Fine-Tuning of Robust Classifiers

Francesco Croce Matthias Hein

University of Tübingen

International Conference on Machine Learning 2022

Some context

- A classifier $f : [0,1]^d \to \mathbb{R}^K$ is robust wrt a **single** l_p-norm at radius ϵ at a point x with correct label c if

$$\arg\max_{r=1,\dots,K} f_r(x+\delta) = c, \quad \text{for every } \delta \quad \text{s. th. } \|\delta\|_p \leq \epsilon,\ x+\delta \in [0,1]^d$$

- **Adversarial training** is commonly used to obtain robust models $\rightarrow$ **more expensive** than standard training

- **Multiple norm robustness** means simultaneous robustness to several threat models, in our case l_∞, l_2 and l_1

- SOTA methods for multiple norm robustness perform adversarial training for every l_p $\rightarrow$ mostly **more expensive** than adversarial training wrt single norms

Fine-tuning robust classifiers

Goal: obtaining models with multiple norm robustness **efficiently**

Idea: **short fine-tuning** of l_p-robust classifiers for multiple norm robustness

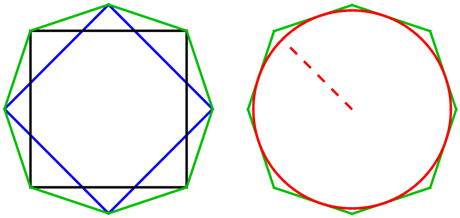
model	*clean*	l_∞ ($\epsilon_\infty = \frac{8}{255}$)	l_2 ($\epsilon_2 = 0.5$)	l_1 ($\epsilon_1 = 12$)	*average*	*union*	*time/epoch*
RN-18 l_∞-AT	83.7	48.1	59.8	7.7	38.5	7.7	151 s
+ SAT	83.5 ± 0.2	43.5 ± 0.2	68.0 ± 0.4	47.4 ± 0.5	53.0 ± 0.2	41.0 ± 0.3	161 s
+ AVG	84.2 ± 0.4	43.3 ± 0.4	68.4 ± 0.6	46.9 ± 0.6	52.9 ± 0.4	40.6 ± 0.4	479 s
+ MAX	82.2 ± 0.3	45.2 ± 0.4	67.0 ± 0.7	46.1 ± 0.4	52.8 ± 0.3	42.2 ± 0.6	466 s
+ MSD	82.2 ± 0.4	44.9 ± 0.3	67.1 ± 0.6	47.2 ± 0.6	53.0 ± 0.4	42.6 ± 0.2	306 s
+ E-AT	82.7 ± 0.4	44.3 ± 0.6	68.1 ± 0.5	48.7 ± 0.5	53.7 ± 0.3	42.2 ± 0.8	160 s

Fine-tuning l_p-robust models with any $p \in \{\infty, 2, 1\}$ for multiple norm robustness for **3 epochs (CIFAR-10) or 1 epoch (ImageNet)** is sufficient to reach competitive robustness in the union of threat models!

Extreme norm Adversarial Training

Problem: MAX (Tramèr & Boneh, 2019) and MSD (Maini et al., 2020) are 2-3x **more expensive** than single norm adversarial training.

Note: Croce & Hein (2020) show that, for linear classifiers, robustness wrt l_∞ and l_1 (extreme norms) is sufficient for robustness wrt l_p for $p \in (1, \infty)$.



We propose **Extreme norm Adversarial Training (E-AT)**, which

- performs adversarial training for a **single** norm, l_∞ or l_1, for each batch,
- **adaptively** samples the threat model to use,
- is **as expensive** as single norm adversarial training.

CIFAR-10 — Fine-tuning l_∞-robust models

model		clean		l_∞ ($\epsilon_\infty = \frac{8}{255}$)		l_2 ($\epsilon_2 = 0.5$)		l_1 ($\epsilon_1 = 12$)		union	
RN-50 - l_∞		88.7		50.9		59.4		5.0		5.0	
(Engstrom et al., 2019)	+ FT	86.2	-2.5	46.0	-4.9	70.1	10.7	49.2	44.2	43.4	38.4
WRN-34-20 - l_∞		87.2		56.6		63.7		8.5		8.5	
(Gowal et al., 2020)	+ FT	88.3	1.1	49.3	-7.3	71.8	8.1	51.2	42.7	46.2	37.7
WRN-28-10 - l_∞ (*)		90.3		59.1		65.7		8.0		8.0	
(Carmon et al., 2019)	+ FT	90.3	0.0	52.6	-6.5	74.7	9.0	54.0	46.0	48.7	40.7
WRN-28-10 - l_∞ (*)		89.9		62.9		67.2		10.8		10.8	
(Gowal et al., 2020)	+ FT	91.2	1.3	53.9	-9.0	76.0	8.8	56.9	46.1	50.1	39.3
WRN-70-16 - l_∞ (*)		90.7		65.6		66.9		8.1		8.1	
(Gowal et al., 2020)	+ FT	91.6	0.9	54.3	-11.3	78.2	11.3	58.3	50.2	51.2	43.1

ImageNet — Fine-tuning l_∞-robust models

model		clean		l_∞ ($\epsilon_\infty = \frac{4}{255}$)		l_2 ($\epsilon_2 = 2$)		l_1 ($\epsilon_1 = 255$)		union	
RN-50 - l_∞		62.9		29.8		17.7		0.0		0.0	
(Engstrom et al., 2019)	+ FT	58.0	-4.9	27.3	-2.5	41.1	23.4	24.0	24.0	21.7	21.7
RN-50 - l_∞		68.2		36.7		15.6		0.0		0.0	
(Bai et al., 2021)	+ FT	60.1	-8.1	29.2	-7.5	42.1	26.5	24.5	24.5	22.6	22.6
DeiT-S - l_∞		66.4		35.6		40.1		3.1		3.1	
(Bai et al., 2021)	+ FT	62.6	-3.8	32.2	-3.4	46.1	6.0	24.8	21.7	23.6	20.5
XCiT-S - l_∞		72.8		41.7		45.3		2.7		2.7	
(Debenedetti, 2022)	+ FT	68.0	-4.8	36.4	-5.3	51.3	6.0	28.4	25.7	26.7	24.0

Quick fine-tuning with E-AT is effective on different architectures, datasets, with or without extra data.

CIFAR-10

Fine-tuning l_∞-robust models

model		clean		l_∞ ($\epsilon_\infty = \frac{8}{255}$)		l_2 ($\epsilon_2 = 0.5$)		l_1 ($\epsilon_1 = 12$)		union	
RN-50 - l_∞ (Engstrom et al., 2019)		88.7		50.9		59.4		5.0		5.0	
	+ FT	86.2	-2.5	46.0	-4.9	70.1	10.7	49.2	44.2	43.4	38.4
WRN-34-20 - l_∞ (Gowal et al., 2020)		87.2		56.6		63.7		8.5		8.5	
	+ FT	88.3	1.1	49.3	-7.3	71.8	8.1	51.2	42.7	46.2	37.7
WRN-28-10 - l_∞ (*) (Carmon et al., 2019)		90.3		59.1		65.7		8.0		8.0	
	+ FT	90.3	0.0	52.6	-6.5	74.7	9.0	54.0	46.0	48.7	40.7
WRN-28-10 - l_∞ (*) (Gowal et al., 2020)		89.9		62.9		67.2		10.8		10.8	
	+ FT	91.2	1.3	53.9	-9.0	76.0	8.8	56.9	46.1	50.1	39.3
WRN-70-16 - l_∞ (*) (Gowal et al., 2020)		90.7		65.6		66.9		8.1		8.1	
	+ FT	91.6	0.9	54.3	-11.3	78.2	11.3	58.3	50.2	51.2	43.1

ImageNet

Fine-tuning l_∞-robust models

model		clean		l_∞ ($\epsilon_\infty = \frac{4}{255}$)		l_2 ($\epsilon_2 = 2$)		l_1 ($\epsilon_1 = 255$)		union	
RN-50 - l_∞ (Engstrom et al., 2019)		62.9		29.8		17.7		0.0		0.0	
	+ FT	58.0	-4.9	27.3	-2.5	41.1	23.4	24.0	24.0	21.7	21.7
RN-50 - l_∞ (Bai et al., 2021)		68.2		36.7		15.6		0.0		0.0	
	+ FT	60.1	-8.1	29.2	-7.5	42.1	26.5	24.5	24.5	22.6	22.6
DeiT-S - l_∞ (Bai et al., 2021)		66.4		35.6		40.1		3.1		3.1	
	+ FT	62.6	-3.8	32.2	-3.4	46.1	6.0	24.8	21.7	23.6	20.5
XCiT-S - l_∞ (Debenedetti, 2022)		72.8		41.7		45.3		2.7		2.7	
	+ FT	68.0	-4.8	36.4	-5.3	51.3	6.0	28.4	25.7	26.7	24.0

Quick fine-tuning with E-AT allows to obtain SOTA multiple norm robustness with large architectures or datasets with low computational cost!

Why multiple norm robustness?

We test the robustness of various classifiers on CIFAR-10 to **unseen non l_p-bounded** attacks (sparse attacks, adversarial corruptions).

model	*clean*	*comm. corr.*	l_0	*patches*	*frames*	*fog*	*snow*	*gabor*	*elastic*	*jpeg*	*avg.*	*union*
NAT	94.4	71.6	0.1	8.1	2.6	47.3	3.9	35.0	0.2	0.0	12.2	0.0
l_∞-AT	81.9	72.6	7.3	21.6	26.2	36.0	35.9	52.5	59.4	5.1	30.5	2.0
l_2-AT	87.8	79.2	13.2	25.0	17.7	44.9	22.1	43.5	56.6	14.0	29.6	4.5
l_1-AT	83.5	75.0	40.9	41.3	21.1	35.6	20.6	41.2	53.3	25.5	34.9	8.6
PAT	82.6	76.9	23.3	37.9	21.7	53.5	25.6	41.8	53.5	13.7	33.9	8.0
SAT	80.5	72.0	38.7	36.7	29.3	33.5	29.0	49.8	57.0	37.4	38.9	13.8
AVG	82.0	73.6	39.7	36.8	30.8	37.2	21.1	49.9	58.1	30.4	38.0	10.9
MAX	80.1	71.3	35.1	34.6	32.7	34.5	35.0	53.4	58.5	33.5	39.7	15.3
MSD	81.0	71.7	36.9	35.0	31.8	34.6	26.4	51.5	59.7	33.4	38.7	12.9
E-AT	79.1	71.3	39.5	37.7	30.5	34.8	33.4	50.2	58.6	38.7	40.4	15.9

Models trained wrt multiple norms show the highest robustness to unseen attacks.

Fine-tuning to another l_q-threat model

We try to fine-tune a classifier robust wrt l_p with adversarial training wrt l_q for $q \neq p$ (3 epochs for CIFAR-10, 1 epoch for ImageNet).

CIFAR-10

	clean	l_∞	l_2	l_1
WRN-70-16 (Gowal et al., 2020) - l_∞ (*)				
original	90.7	65.6	66.9	8.1
+ FT wrt l_2	92.8	47.4	80.0	34.0
+ FT wrt l_1	92.4	33.9	74.7	**70.2**
WRN-70-16 (Gowal et al., 2020) - l_2 (*)				
original	94.1	43.1	81.7	34.6
+ FT wrt l_∞	92.3	58.5	73.5	11.4
+ FT wrt l_1	92.8	29.2	75.7	68.9
RN-18 (Croce & Hein, 2021) - l_1				
original	87.1	22.0	64.8	60.3
+ FT wrt l_∞	82.7	44.2	66.6	25.4
+ FT wrt l_2	88.0	31.0	69.8	39.7

ImageNet

	clean	l_∞	l_2	l_1
DeiT-S (Bai et al., 2021) - l_∞				
original	66.4	35.6	40.1	3.1
+ FT wrt l_2	66.5	31.2	46.1	9.6
+ FT wrt l_1	61.0	23.9	42.9	30.1
XCiT-S (Debenedetti, 2022) - l_∞				
original	72.8	41.7	45.3	2.7
+ FT wrt l_2	71.5	35.9	51.4	9.5
+ FT wrt l_1	65.8	25.2	47.1	**33.9**
RN-50 (Engstrom et al., 2019) - l_2				
original	58.7	25.0	40.5	14.0
+ FT wrt l_∞	59.1	31.5	40.1	7.5
+ FT wrt l_1	56.8	18.0	37.1	28.7

Fine-tuning robust classifiers allows to quickly obtain competitive baselines in other threat models!