Making and **Measuring** Progress in Adversarial Machine Learning

Nicholas Carlini
Google Research

Act I
Background

88% **tabby cat**

adversarial
perturbation

88% **tabby cat**

adversarial perturbation

88% **tabby cat**

88% tabby cat

adversarial perturbation

99% guacamole

Why should we care about adversarial examples?

Make ML ***robust***

Make ML ***better***

Act II
An Apparent Problem

Let's go back
to ~5 years ago ...

Generative Adversarial Nets

SotA, 2014

Progressive Growing of GANs

SotA, 2017

Evasion Attacks against ML at Test Time

SotA, 2013

Exploiting Excessive Invariance caused by Norm-Bounded Adversarial Robustness

SotA, 2019

that is ...
... less impressive

3 years:  

6 years:  

Why?

Act III

Measuring Progress

Have we even made ***any*** progress?

A Brief History of ~~time~~ defenses

- Oakland'16 - *broken*
- ICLR'17 - *broken*
- CCS'17 - *broken*
- ICLR'18 - *broken (mostly)*
- CVPR'18 - *broken*
- NeurIPS'18 - *broken (some)*

Have we even made
any progress?

Is this a constant cat-and-mouse game?

What does it mean to make progress?

What does it mean to make progress?

Learning something **new**.

A Brief History of ~~time~~ defenses

- Oakland'16 - *gradient masking*
- ICLR'17 - *attack objective functions*
- CCS'17 - *transferability of examples*
- ICLR'18 - *obfuscated gradients*

A Brief History of ~~time~~ defenses

- Oakland'16 - *gradient masking*
- ICLR'17 - *attack objective functions*
- CCS'17 - *transferability of examples*
- ICLR'18 - *obfuscated gradients*
- 2019 - **???**

Measure by how much we learn; not by how much robustness we gain.

Act IV

Making Progress
(for defenses)

While we have learned a lot, it's less than I would have hoped.

Cargo Cult Evaluations

Going through the motions is

insufficient

to do proper security evaluations

An all too common paper:

3.1. Effectiveness

3.1. Effectiveness

Adversarial Attacks. We test on the following attacks:

we trained on and L_{CW} is an objective encouraging misclassification. Under this threat model, *NeuralFP* achieves an AUC-ROC of **98.79%** against Adaptive-CW-L_2, with $N = 30$ and $\epsilon = 0.006$ for a set of unseen test-samples (1024 *pre-test*) and the corresponding adversarial examples. In contrast to other defenses that are vulnerable to Adaptive-CW-L_2 (Carlini & Wagner, 2017a), we find that *NeuralFP* is robust even under this whitebox-attack threat model.

4. Related Work

5. Discussion and Future Work

3.4. Robustness to Adaptive Whitebox-Attackers

We further considered an adaptive attacker that has knowledge of the predetermined fingerprints and model weights, similar to (Carlini & Wagner, 2017a). Here, the adaptive attacker (Adaptive-CW-L_2) tries to find an adversarial example x' that also minimizes the fingerprint-loss, attacking a CIFAR-10 model trained with *NeuralFP*. To this end, the CW-L_2 objective is modified as:

$$\min_{x'} \|x - x'\|_2 + \gamma \left(L_{CW}(x') + L_{fp}(x', y^*, \xi; \theta)\right) \quad (29)$$

Here, y^* is the label-vector, $\gamma \in [10^{-3}, 10^6]$ is a scalar found through a bisection search, L_{fp} is the fingerprint-loss

An all too common paper:

The two types of defenses:

Defenses that are broken by **existing attacks**

Defenses that are broken by **new attacks**

Exciting new directions

Exciting new directions

SentiNet: Detecting Physical Attacks Against Deep Learning Systems

Edward Chou[1] Florian Tramèr[1] Giancarlo Pellegrino[1,2] Dan Boneh[1]

Exciting new directions

Sitatapatra: Blocking the Transfer of Adversarial Samples

Ilia Shumailov [*1] Xitong Gao [*2] Yiren Zhao [*1] Robert Mullins [1] Ross Anderson [1] Cheng-Zhong Xu [2]

Exciting new directions

Adversarial Examples Are Not Bugs, They Are Features

Andrew Ilyas*
MIT
ailyas@mit.edu

Shibani Santurkar*
MIT
shibani@mit.edu

Dimitris Tsipras*
MIT
tsipras@mit.edu

Logan Engstrom*
MIT
engstrom@mit.edu

Brandon Tran
MIT
btran115@mit.edu

Aleksander Mądry
MIT
madry@mit.edu

Act IV ½

Making Progress
(for attacks)

Advice for performing evaluations

ON EVALUATING ADVERSARIAL ROBUSTNESS

Nicholas Carlini[1], Anish Athalye[2], Nicolas Papernot[1], Wieland Brendel[3], Jonas Rauber[3], Dimitris Tsipras[2], Ian Goodfellow[1], Aleksander Mądry[2], Alexey Kurakin[1] *

[1] Google Brain [2] MIT [3] University of Tübingen

Perform Adaptive Attacks

An all too common paper:

CLASSIFICATION ACCURACY OF COMPLETELY DEFENSES AGAINST ADVERSARIAL ATTACKS ON MNIST

Datasets	UA/TA	Objective	Attacks		# of AEs	Original Model	NAT	EAT	PAT	DD	IGR	EIT	RT	PD	TE	RC	Average
							Adversarial Training			Gradient Masking		Input Transformation					
MNIST	UAs	L_∞ $\epsilon=0.3$	FGSM	$\epsilon=0.3$	304	0.0%	88.2%	88.8%	94.1%	60.2%	76.6%	61.5%	26.0%	11.2%	93.8%	8.6%	60.9%
				$\epsilon=0.5$	448	0.0%	28.8%	25.9%	26.3%	29.5%	42.6%	27.9%	15.2%	2.9%	20.5%	1.6%	22.1%
			R+FGSM		342	0.0%	95.9%	95.6%	98.3%	78.4%	88.3%	77.8%	30.1%	17.3%	97.1%	19.3%	69.8%
			BIM		756	0.0%	93.0%	92.5%	97.9%	71.6%	83.6%	65.7%	21.2%	7.0%	97.6%	9.4%	63.9%
			PGD		824	0.0%	95.4%	93.8%	98.2%	74.5%	85.9%	67.7%	18.3%	10.6%	98.4%	11.5%	65.4%
			U-MI-FGSM		704	0.0%	90.9%	90.8%	97.3%	64.4%	80.0%	57.0%	20.2%	8.7%	97.0%	9.8%	61.6%
			UAP		303	0.0%	96.7%	95.4%	98.7%	71.3%	15.2%	12.9%	10.2%	18.2%	97.7%	42.2%	55.8%
		L_2	DF		1000	0.0%	99.6%	99.3%	99.3%	96.8%	98.9%	99.0%	68.9%	97.5%	99.3%	98.6%	95.7%
			OM		1000	0.0%	88.6%	87.7%	93.7%	70.8%	90.5%	77.9%	26.3%	4.8%	94.8%	1.0%	63.6%
	TAs	L_∞ $\epsilon=0.3$	LLC		56	0.0%	96.4%	98.2%	100.0%	73.2%	[illegible]	[illegible]	[illegible]	[illegible]	100.0%	7.1%	67.9%
			R+LLC		40	0.0%	97.5%	95.0%	97.5%	92.5%	[illegible]	[illegible]	[illegible]	[illegible]	95.0%	32.5%	75.5%
			ILLC		594	0.0%	98.7%	98.8%	99.0%	87.0%	[illegible]	[illegible]	[illegible]	[illegible]	98.5%	30.8%	74.1%
			T-MI-FGSM		864	0.0%	98.4%	97.9%	99.2%	81.9%	[illegible]	[illegible]	[illegible]	[illegible]	99.1%	22.9%	70.7%
		L_0	JSMA		764	0.0%	78.5%	74.1%	79.7%	78.5%	86.0%	73.7%	38.6%	13.7%	73.0%	35.2%	63.1%
		L_2	BLB		1000	0.0%	99.7%	99.0%	99.3%	98.7%	99.1%	99.0%	66.3%	95.2%	98.6%	98.0%	95.3%
			CW2	$\kappa=0$	997	0.0%	99.6%	99.0%	99.3%	98.2%	99.1%	98.2%	68.3%	96.4%	98.4%	98.4%	95.5%
				$\kappa=20$	963	0.0%	79.7%	78.7%	85.2%	79.2%	90.7%	73.5%	19.8%	2.9%	81.1%	0.5%	59.1%
			EAD	EN	1000	0.0%	99.3%	98.4%	99.0%	98.4%	99.1%	98.5%	64.5%	96.6%	98.0%	98.3%	95.0%
				L1	1000	0.0%	99.0%	97.8%	98.3%	98.1%	99.1%	97.9%	62.0%	94.9%	96.8%	98.3%	94.2%
	Average				682.1	0.0%	90.7%	89.8%	92.6%	79.1%	85.0%	74.4%	33.8%	35.3%	91.3%	38.1%	**71.0%**

98.2%

Ensure correct implementations

An all too common paper:

CLASSIFICATION ACCURACY OF COMPLETELY DEFENSES AGAINST ADVERSARIAL ATTACKS ON MNIST

Datasets	Attack					Original Model	Defense-enhanced Models										Average
	UA/TA	Objective	Attacks		# of AEs		Adversarial Training			Gradient Masking		Input Transformation				RC	
							NAT	EAT	PAT	DD	IGR	EIT	RT	PD	TE		
MNIST	UAs	L_∞ $\epsilon=0.3$	FGSM	$\epsilon=0.3$	304	0.0%	88.2%	88.8%	94.1%	60.2%	76.6%	61.5%	26.0%	11.2%	93.8%	8.6%	60.9%
				$\epsilon=0.5$	448	0.0%	28.8%	25.9%	26.3%	29.5%	42.6%	27.9%	15.2%	2.9%	20.5%	1.6%	22.1%
			R+FGSM		342	0.0%	95.9%	95.6%	98.3%	78.4%	88.3%	77.8%	30.1%	17.3%	97.1%	19.3%	69.8%
			BIM		756	0.0%	93.0%	92.5%	97.9%	71.6%	83.6%	65.7%	21.2%	7.0%	97.6%	9.4%	63.9%
			PGD		824	0.0%	95.4%	93.8%	98.2%	74.5%	85.9%	67.7%	18.3%	10.6%	98.4%	11.5%	65.4%
			U-MI-FGSM		704	0.0%	90.9%	90.8%	97.3%	64.4%	80.0%	57.0%	20.2%	8.7%	97.0%	9.8%	61.6%
			UAP		303	0.0%	96.7%	95.4%	98.7%	71.3%	15.2%	12.9%	10.2%	18.2%	97.7%	42.2%	55.8%
		L_2	DF		1000	0.0%	99.6%	99.3%	99.3%	96.8%	98.9%	99.0%	68.9%	97.5%	99.3%	98.6%	95.7%
			OM		1000	0.0%	88.6%	87.7%	93.7%	70.8%	90.5%	77.9%	26.3%	4.8%	94.8%	1.0%	63.6%
	TAs	L_∞ $\epsilon=0.3$	LLC		[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	23.2%	100.0%	7.1%	67.9%
			R+LLC		[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	25.0%	30.0%	95.0%	32.5%	75.5%
			ILLC		[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	25.4%	20.7%	98.5%	30.8%	74.1%
			T-MI-FGSM		864	0.0%	98.4%	97.9%	99.2%	81.9%	90.5%	72.1%	26.6%	18.8%	99.1%	22.9%	70.7%
		L_0	JSMA		764	0.0%	78.5%	74.1%	79.7%	78.5%	86.0%	73.7%	38.6%	13.7%	73.0%	35.2%	63.1%
		L_2	BLB		1000	0.0%	99.7%	99.0%	99.3%	98.7%	99.1%	99.0%	66.3%	95.2%	98.6%	98.0%	95.3%
			CW2	$\kappa=0$	997	0.0%	99.6%	99.0%	99.3%	98.2%	99.1%	98.2%	68.3%	96.4%	98.4%	98.4%	95.5%
				$\kappa=20$	963	0.0%	79.7%	78.7%	85.2%	79.2%	90.7%	73.5%	19.8%	2.9%	81.1%	0.5%	59.1%
			EAD	EN	1000	0.0%	99.3%	98.4%	99.0%	98.4%	99.1%	98.5%	64.5%	96.6%	98.0%	98.3%	95.0%
				L1	1000	0.0%	99.0%	97.8%	98.3%	98.1%	99.1%	97.9%	62.0%	94.9%	96.8%	98.3%	94.2%
	Average				682.1	0.0%	90.7%	89.8%	92.6%	79.1%	85.0%	74.4%	33.8%	35.3%	91.3%	38.1%	**71.0%**

JSMA	764

An all too common paper:

CLASSIFICATION ACCURACY OF COMPLETELY DEFENSES AGAINST ADVERSARIAL ATTACKS ON MNIST

Datasets	UA/TA	Objective	Attacks	# of AEs	Original Model	Adversarial Training: NAT	Adversarial Training: EAT	Adversarial Training: PAT	Gradient Masking: DD	Gradient Masking: IGR	Input Transformation: EIT	Input Transformation: RT	Input Transformation: PD	Input Transformation: TE	RC	Average
MNIST	UAs	L_∞ $\epsilon=0.3$	FGSM $\epsilon=0.3$	304									11.2%	93.8%	8.6%	60.9%
			FGSM $\epsilon=0.5$	448									2.9%	20.5%	1.6%	22.1%
			R+FGSM	342									17.3%	97.1%	19.3%	69.8%
			BIM	756									7.0%	97.6%	9.4%	63.9%
			PGD	824									10.6%	98.4%	11.5%	65.4%
			U-MI-FGSM	704	0.0%	90.9%	90.8%	97.3%	64.4%	80.0%	57.0%	20.2%	8.7%	97.0%	9.8%	61.6%
			UAP	303	0.0%	96.7%	95.4%	98.7%	71.3%	15.2%	12.9%	10.2%	18.2%	97.7%	42.2%	55.8%
		L_2	DF	1000	0.0%	99.6%	99.3%	99.3%	96.8%	98.9%	99.0%	68.9%	97.5%	99.3%	98.6%	95.7%
			OM	1000	0.0%	88.6%	87.7%	93.7%	70.8%	90.5%	77.9%	26.3%	4.8%	94.8%	1.0%	63.6%
	TAs	L_∞ $\epsilon=0.3$	LLC	56	0.0%	96.4%	98.2%	100.0%	73.2%	96.4%	75.0%	8.9%	23.2%	100.0%	7.1%	67.9%
			R+LLC	40	0.0%	97.5%	95.0%	97.5%	92.5%	97.5%	92.5%	25.0%	30.0%	95.0%	32.5%	75.5%
			ILLC	594	0.0%	98.7%	98.8%	99.0%	87.0%	95.1%	86.7%	25.4%	20.7%	98.5%	30.8%	74.1%
			T-MI-FGSM	864	0.0%	98.4%	97.9%	99.2%	81.9%	90.5%	72.1%	26.6%	18.8%	99.1%	22.9%	70.7%
		L_0	JSMA	764	0.0%	78.5%	74.1%	79.7%	78.5%	86.0%	73.7%	38.6%	13.7%	73.0%	35.2%	63.1%
		L_2	BLB	1000	0.0%	99.7%	99.0%	99.3%	98.7%	99.1%	99.0%	66.3%	95.2%	98.6%	98.0%	95.3%
			CW2 $\kappa=0$	997	0.0%	99.6%	99.0%	99.3%	98.2%	99.1%	98.2%	68.3%	96.4%	98.4%	98.4%	95.5%
			CW2 $\kappa=20$	963	0.0%	79.7%	78.7%	85.2%	79.2%	90.7%	73.5%	19.8%	2.9%	81.1%	0.5%	59.1%
			EAD EN	1000	0.0%	99.3%	98.4%	99.0%	98.4%	99.1%	98.5%	64.5%	96.6%	98.0%	98.3%	95.0%
			EAD L1	1000	0.0%	99.0%	97.8%	98.3%	98.1%	99.1%	97.9%	62.0%	94.9%	96.8%	98.3%	94.2%
Average				682.1	0.0%	90.7%	89.8%	92.6%	79.1%	85.0%	74.4%	33.8%	35.3%	91.3%	38.1%	**71.0%**

BIM	756
PGD	824

Use meaningful threat models

An all too common paper:

CLASSIFICATION ACCURACY OF COMPLETELY DEFENSES AGAINST ADVERSARIAL ATTACKS ON MNIST

Datasets	Attack					Original Model	Defense-enhanced Models										Average
	UA/ TA	Objec- tive	Attacks		# of AEs		Adversarial Training			Gradient Masking		Input Transformation				RC	
							NAT	EAT	PAT	DD	IGR	EIT	RT	PD	TE		
MNIST	UAs	L_∞ $\epsilon=0.3$	FGSM	$\epsilon=0.3$	304	0.0%	88.2%	88.8%	94.1%	60.2%	76.6%	61.5%	26.0%	11.2%	93.8%	8.6%	60.9%
				$\epsilon=0.5$	448	0.0%	28.8%	25.9%	26.3%	29.5%	42.6%	27.9%	15.2%	2.9%	20.5%	1.6%	22.1%
			R+FGSM		342	0.0%	95.9%	95.6%	98.3%	78.4%	88.3%	77.8%	30.1%	17.3%	97.1%	19.3%	69.8%
			BIM		[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	83.6%	65.7%	21.2%	7.0%	97.6%	9.4%	63.9%
			PGD		[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	85.9%	67.7%	18.3%	10.6%	98.4%	11.5%	65.4%
			U-MI-FGSM		[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	80.0%	57.0%	20.2%	8.7%	97.0%	9.8%	61.6%
			UAP		[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	15.2%	12.9%	10.2%	18.2%	97.7%	42.2%	55.8%
		L_2	DF		[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	98.9%	99.0%	68.9%	97.5%	99.3%	98.6%	95.7%
			OM		[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	90.5%	77.9%	26.3%	4.8%	94.8%	1.0%	63.6%
	TAs	L_∞ $\epsilon=0.3$	LLC		56	0.0%	96.4%	98.2%	100.0%	73.2%	96.4%	75.0%	8.9%	23.2%	100.0%	7.1%	67.9%
			R+LLC		40	0.0%	97.5%	95.0%	97.5%	92.5%	97.5%	92.5%	25.0%	30.0%	95.0%	32.5%	75.5%
			ILLC		594	0.0%	98.7%	98.8%	99.0%	87.0%	95.1%	86.7%	25.4%	20.7%	98.5%	30.8%	74.1%
			T-MI-FGSM		864	0.0%	98.4%	97.9%	99.2%	81.9%	90.5%	72.1%	26.6%	18.8%	99.1%	22.9%	70.7%
		L_0	JSMA		764	0.0%	78.5%	74.1%	79.7%	78.5%	86.0%	73.7%	38.6%	13.7%	73.0%	35.2%	63.1%
		L_2	BLB		1000	0.0%	99.7%	99.0%	99.3%	98.7%	99.1%	99.0%	66.3%	95.2%	98.6%	98.0%	95.3%
			CW2	$\kappa=0$	997	0.0%	99.6%	99.0%	99.3%	98.2%	99.1%	98.2%	68.3%	96.4%	98.4%	98.4%	95.5%
				$\kappa=20$	963	0.0%	79.7%	78.7%	85.2%	79.2%	90.7%	73.5%	19.8%	2.9%	81.1%	0.5%	59.1%
			EAD	EN	1000	0.0%	99.3%	98.4%	99.0%	98.4%	99.1%	98.5%	64.5%	96.6%	98.0%	98.3%	95.0%
				L1	1000	0.0%	99.0%	97.8%	98.3%	98.1%	99.1%	97.9%	62.0%	94.9%	96.8%	98.3%	94.2%
	Average				682.1	0.0%	90.7%	89.8%	92.6%	79.1%	85.0%	74.4%	33.8%	35.3%	91.3%	38.1%	**71.0%**

FGSM $\epsilon=0.3$ $\epsilon=0.5$

An all too common paper:

CLASSIFICATION ACCURACY OF COMPLETELY DEFENSES AGAINST ADVERSARIAL ATTACKS ON MNIST

Datasets	UA/TA	Objective	Attacks		# of AEs	Original Model	NAT	EAT	PAT	DD	IGR	EIT	RT	PD	TE	RC	Average
							Adversarial Training			Gradient Masking		Input Transformation					
MNIST	UAs	L_∞ $\epsilon=0.3$	FGSM	$\epsilon=0.3$	304	0.0%	88.2%	88.8%	94.1%	60.2%	76.6%	61.5%	26.0%	11.2%	93.8%	8.6%	60.9%
				$\epsilon=0.5$	448	0.0%	28.8%	25.9%	26.3%	29.5%	42.6%	27.9%	15.2%	2.9%	20.5%	1.6%	22.1%
			R+FGSM		342	0.0%	95.9%	95.6%	98.3%	78.4%	88.3%	77.8%	30.1%	17.3%	97.1%	19.3%	69.8%
			BIM		[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	21.2%	7.0%	97.6%	9.4%	63.9%
			PGD		[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	18.3%	10.6%	98.4%	11.5%	65.4%
			U-MI-FGSM		[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	20.2%	8.7%	97.0%	9.8%	61.6%
			UAP		[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	10.2%	18.2%	97.7%	42.2%	55.8%
		L_2	DF		[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	68.9%	97.5%	99.3%	98.6%	95.7%
			OM		[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	26.3%	4.8%	94.8%	1.0%	63.6%
	TAs	L_∞ $\epsilon=0.3$	LLC		56	0.0%	96.4%	98.2%	100.0%	73.2%	96.4%	75.0%	8.9%	23.2%	100.0%	7.1%	67.9%
			R+LLC		40	0.0%	97.5%	95.0%	97.5%	92.5%	97.5%	92.5%	25.0%	30.0%	95.0%	32.5%	75.5%
			ILLC		594	0.0%	98.7%	98.8%	99.0%	87.0%	95.1%	86.7%	25.4%	20.7%	98.5%	30.8%	74.1%
			T-MI-FGSM		864	0.0%	98.4%	97.9%	99.2%	81.9%	90.5%	72.1%	26.6%	18.8%	99.1%	22.9%	70.7%
		L_0	JSMA		764	0.0%	78.5%	74.1%	79.7%	78.5%	86.0%	73.7%	38.6%	13.7%	73.0%	35.2%	63.1%
		L_2	BLB		1000	0.0%	99.7%	99.0%	99.3%	98.7%	99.1%	99.0%	66.3%	95.2%	98.6%	98.0%	95.3%
			CW2	$\kappa=0$	997	0.0%	99.6%	99.0%	99.3%	98.2%	99.1%	98.2%	68.3%	96.4%	98.4%	98.4%	95.5%
				$\kappa=20$	963	0.0%	79.7%	78.7%	85.2%	79.2%	90.7%	73.5%	19.8%	2.9%	81.1%	0.5%	59.1%
			EAD	EN	1000	0.0%	99.3%	98.4%	99.0%	98.4%	99.1%	98.5%	64.5%	96.6%	98.0%	98.3%	95.0%
				L1	1000	0.0%	99.0%	97.8%	98.3%	98.1%	99.1%	97.9%	62.0%	94.9%	96.8%	98.3%	94.2%
Average					682.1	0.0%	90.7%	89.8%	92.6%	79.1%	85.0%	74.4%	33.8%	35.3%	91.3%	38.1%	**71.0%**

FGSM $\epsilon=0.3$ 304
$\epsilon=0.5$ 448

An all too common paper:

CLASSIFICATION ACCURACY OF COMPLETELY DEFENSES AGAINST ADVERSARIAL ATTACKS ON MNIST

Datasets	Attack					Original Model	Defense-enhanced Models										Average
	UA/ TA	Objec-tive	Attacks		# of AEs		Adversarial Training			Gradient Masking		Input Transformation				RC	
							NAT	EAT	PAT	DD	IGR	EIT	RT	PD	TE		
MNIST	UAs	L_∞ $\epsilon=0.3$	FGSM	$\epsilon=0.3$	304	0.0%	88.2%	88.8%	94.1%	60.2%	76.6%	61.5%	26.0%	11.2%	93.8%	8.6%	60.9%
				$\epsilon=0.5$	448	0.0%	28.8%	25.9%	26.3%	29.5%	42.6%	27.9%	15.2%	2.9%	20.5%	1.6%	22.1%
			R+FGSM		342	0.0%	95.9%	95.6%	98.3%	78.4%	88.3%	77.8%	30.1%	17.3%	97.1%	19.3%	69.8%
			BIM		756	0.0%	93.0%	92.5%	97.9%	71.6%	83.6%	65.7%	21.2%	7.0%	97.6%	9.4%	63.9%
			PGD		824	0.0%	95.4%	93.8%	98.2%	74.5%	85.9%	67.7%	18.3%	10.6%	98.4%	11.5%	65.4%
			U-MI-FGSM		704	0.0%	90.9%	90.8%	97.3%	64.4%	80.0%	57.0%	20.2%	8.7%	97.0%	9.8%	61.6%
			UAP		303	0.0%	96.7%	95.4%	98.7%	71.3%	15.2%	12.9%	10.2%	18.2%	97.7%	42.2%	55.8%
		L_2	DF		1000	0.0%	99.6%	99.3%	99.3%	96.8%	98.9%	99.0%	68.9%	97.5%	99.3%	98.6%	95.7%
			OM		1000	0.0%	88.6%	87.7%	93.7%	70.8%	90.5%	77.9%	26.3%	4.8%	94.8%	1.0%	63.6%
	TAs	L_∞ $\epsilon=0.3$	LLC		56	0.0%	96.4%	98.2%	100.0%	73.2%	96.4%	75.0%	8.9%	23.2%	100.0%	7.1%	67.9%
			R+LLC		40	0.0%	97.5%	95.0%	97.5%	92.5%	97.5%	92.5%	25.0%	30.0%	95.0%	32.5%	75.5%
			ILLC		594	0.0%	98.7%	98.8%	99.0%	87.0%	95.1%	86.7%	25.4%	20.7%	98.5%	30.8%	74.1%
			T-MI-FGSM		864	0.0%	98.4%	97.9%	99.2%	81.9%	90.5%	72.1%	26.6%	18.8%	99.1%	22.9%	70.7%
		L_0	JSMA		[illegible]	[illegible]	[illegible]	[illegible]	79.7%	78.5%	86.0%	73.7%	38.6%	13.7%	73.0%	35.2%	63.1%
		L_2	BLB		[illegible]	[illegible]	[illegible]	[illegible]	99.3%	98.7%	99.1%	99.0%	66.3%	95.2%	98.6%	98.0%	95.3%
			CW2	$\kappa=0$	[illegible]	[illegible]	[illegible]	[illegible]	99.3%	98.2%	99.1%	98.2%	68.3%	96.4%	98.4%	98.4%	95.5%
				$\kappa=20$	[illegible]	[illegible]	[illegible]	[illegible]	85.2%	79.2%	90.7%	73.5%	19.8%	2.9%	81.1%	0.5%	59.1%
			EAD	EN	[illegible]	[illegible]	[illegible]	[illegible]	99.0%	98.4%	99.1%	98.5%	64.5%	96.6%	98.0%	98.3%	95.0%
				L1	1000	0.0%	99.0%	97.8%	98.3%	98.1%	99.1%	97.9%	62.0%	94.9%	96.8%	98.3%	94.2%
	Average				682.1	0.0%	90.7%	89.8%	92.6%	79.1%	85.0%	74.4%	33.8%	35.3%	91.3%	38.1%	71.0%

L1

Compute Worst-Case Robustness

An all too common paper:

CLASSIFICATION ACCURACY OF COMPLETELY DEFENSES AGAINST ADVERSARIAL ATTACKS ON MNIST

Datasets	Attack					Original Model	Defense-enhanced Models										Average
	UA/TA	Objective	Attacks		# of AEs		Adversarial Training			Gradient Masking		Input Transformation				RC	
							NAT	EAT	PAT	DD	IGR	EIT	RT	PD	TE		
MNIST	UAs	L_∞ $\epsilon=0.3$	FGSM	$\epsilon=0.3$	304	0.0%	88.2%	88.8%	94.1%	60.2%	76.6%	61.5%	26.0%	11.2%	93.8%	8.6%	60.9%
			FGSM	$\epsilon=0.5$	448	0.0%	28.8%	25.9%	26.3%	29.5%	42.6%	27.9%	15.2%	2.9%	20.5%	1.6%	22.1%
			R+FGSM		342	0.0%	95.9%	95.6%	98.3%	78.4%	88.3%	77.8%	30.1%	17.3%	97.1%	19.3%	69.8%
			BIM		756	0.0%	93.0%	92.5%	97.9%	71.6%	83.6%	65.7%	21.2%	7.0%	97.6%	9.4%	63.9%
			PGD		824	0.0%	95.4%	93.8%	98.2%	74.5%	85.9%	67.7%	18.3%	10.6%	98.4%	11.5%	65.4%
			U-MI-FGSM		704	0.0%	90.9%	90.8%	97.3%	64.4%	80.0%	57.0%	20.2%	8.7%	97.0%	9.8%	61.6%
			UAP		303	0.0%	96.7%	95.4%	98.7%	71.3%	15.2%	12.9%	10.2%	18.2%	97.7%	42.2%	55.8%
		L_2	DF		1000	0.0%	99.6%	99.3%	99.3%	96.8%	98.9%	99.0%	68.9%	97.5%	99.3%	98.6%	95.7%
			OM		1000	0.0%	88.6%	87.7%	93.7%	70.8%	90.5%	77.9%	26.3%	4.8%	94.8%	1.0%	63.6%
	TAs	L_∞ $\epsilon=0.3$	LLC		56	0.0%	96.4%	98.2%	100.0%	73.2%	96.4%	75.0%	8.9%	23.2%	100.0%	7.1%	67.9%
			R+LLC		40	0.0%	97.5%	95.0%	97.5%	92.5%	97.5%	92.5%	25.0%	30.0%	95.0%	32.5%	75.5%
			ILLC		594	0.0%	98.7%	98.8%	99.0%	87.0%	95.1%	86.7%	25.4%	20.7%	98.5%	30.8%	74.1%
			T-MI-FGSM		864	0.0%	98.4%	97.9%	99.2%	81.9%	90.5%	72.1%	26.6%	18.8%	99.1%	22.9%	70.7%
		L_0	JSMA		764	0.0%	78.5%	74.1%	79.7%	78.5%	86.0%	73.7%	38.6%	13.7%	73.0%	35.2%	63.1%
		L_2	BLB		1000	0.0%	99.7%	99.0%	99.3%	98.7%	99.1%	99.0%	66.3%	95.2%	98.6%	98.0%	95.3%
			[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	98.2%	99.1%	98.2%	68.3%	96.4%	98.4%	98.4%	95.5%
			[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	79.2%	90.7%	73.5%	19.8%	2.9%	81.1%	0.5%	59.1%
			[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	98.4%	99.1%	98.5%	64.5%	96.6%	98.0%	98.3%	95.0%
			[illegible]	L1	1000	0.0%	99.0%	97.8%	98.3%	98.1%	99.1%	97.9%	62.0%	94.9%	96.8%	98.3%	94.2%
	Average				682.1	0.0%	90.7%	89.8%	92.6%	79.1%	85.0%	74.4%	33.8%	35.3%	91.3%	38.1%	71.0%

Average

An all too common paper:

CLASSIFICATION ACCURACY OF COMPLETELY DEFENSES AGAINST ADVERSARIAL ATTACKS ON MNIST

Datasets	Attack					Original Model	Defense-enhanced Models										Average
	UA/ TA	Objec- tive	Attacks		# of AEs		Adversarial Training			Gradient Masking		Input Transformation				RC	
							NAT	EAT	PAT	DD	IGR	EIT	RT	PD	TE		
MNIST	UAs	L_∞ $\epsilon=0.3$	FGSM	$\epsilon=0.3$	304	0.0%	88.2%	88.8%	94.1%	60.2%	76.6%	61.5%	26.0%	11.2%	93.8%	8.6%	60.9%
				$\epsilon=0.5$	448	0.0%	28.8%	25.9%	26.3%	29.5%	42.6%	27.9%	15.2%	2.9%	20.5%	1.6%	22.1%
			R+FGSM		342	0.0%	95.9%	95.6%	98.3%	78.4%	88.3%	77.8%	30.1%	17.3%	97.1%	19.3%	69.8%
			BIM		756	0.0%	93.0%	92.5%	97.9%	71.6%	83.6%	65.7%	21.2%	7.0%	97.6%	9.4%	63.9%
			PGD		824	0.0%	95.4%	93.8%	98.2%	74.5%	85.9%	67.7%	18.3%	10.6%	98.4%	11.5%	65.4%
			U-MI-FGSM		704	0.0%	90.9%	90.8%	97.3%	64.4%	80.0%	57.0%	20.2%	8.7%	97.0%	9.8%	61.6%
			UAP		303	0.0%	96.7%	95.4%	98.7%	71.3%	15.2%	12.9%	10.2%	18.2%	97.7%	42.2%	55.8%
		L_2	DF		1000	0.0%	99.6%	99.3%	99.3%	96.8%	98.9%	99.0%	68.9%	97.5%	99.3%	98.6%	95.7%
			OM		1000	0.0%	88.6%	87.7%	93.7%	70.8%	90.5%	77.9%	26.3%	4.8%	94.8%	1.0%	63.6%
	TAs	L_∞ $\epsilon=0.3$	LLC		56	0.0%	96.4%	98.2%	100.0%	73.2%	96.4%	75.0%	8.9%	23.2%	100.0%	7.1%	67.9%
			R+LLC		40	0.0%	97.5%	95.0%	97.5%	92.5%	97.5%	92.5%	25.0%	30.0%	95.0%	32.5%	75.5%
			ILLC		594	0.0%	98.7%	98.8%	99.0%	87.0%	95.1%	86.7%	25.4%	20.7%	98.5%	30.8%	74.1%
			T-MI-FGSM		864	0.0%	98.4%	97.9%	99.2%	81.9%	90.5%	72.1%	26.6%	18.8%	99.1%	22.9%	70.7%
		L_0	JSMA		764	0.0%	78.5%	74.1%	79.7%	78.5%	86.0%	73.7%	38.6%	13.7%	73.0%	35.2%	63.1%
		L_2	BLB		1000	0.0%	99.7%	99.0%	99.3%	98.7%	99.1%	99.0%	66.3%	95.2%	98.6%	98.0%	95.3%
			C[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	98.2%	99.1%	98.2%	68.3%	96.4%	98.4%	98.4%	95.5%
			[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	79.2%	90.7%	73.5%	19.8%	2.9%	81.1%	0.5%	59.1%
			E[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	98.4%	99.1%	98.5%	64.5%	96.6%	98.0%	98.3%	95.0%
				L1	1000	0.0%	99.0%	97.8%	98.3%	98.1%	99.1%	97.9%	62.0%	94.9%	96.8%	98.3%	94.2%
	Average				682.1	0.0%	90.7%	89.8%	92.6%	79.1%	85.0%	74.4%	33.8%	35.3%	91.3%	38.1%	**71.0%**

Average

An all too common paper:

CLASSIFICATION ACCURACY OF COMPLETELY DEFENSES AGAINST ADVERSARIAL ATTACKS ON MNIST

Datasets	Attack					Original Model	Defense-enhanced Models										Average
	UA/ TA	Objec- tive	Attacks		# of AEs		Adversarial Training			Gradient Masking		Input Transformation				RC	
							NAT	EAT	PAT	DD	IGR	EIT	RT	PD	TE		
MNIST	UAs	L_∞ $\epsilon=0.3$	FGSM	$\epsilon=0.3$	304	0.0%	88.2%	88.8%	94.1%	60.2%	76.6%	61.5%	26.0%	11.2%	93.8%	8.6%	60.9%
				$\epsilon=0.5$	448	0.0%	28.8%	25.9%	26.3%	29.5%	42.6%	27.9%	15.2%	2.9%	20.5%	1.6%	22.1%
			R+FGSM		342	0.0%	95.9%	95.6%	98.3%	78.4%	88.3%	77.8%	30.1%	17.3%	97.1%	19.3%	69.8%
			BIM		756	0.0%	93.0%	92.5%	97.9%	71.6%	83.6%	65.7%	21.2%	7.0%	97.6%	9.4%	63.9%
			PGD		824	0.0%	95.4%	93.8%	98.2%	74.5%	85.9%	67.7%	18.3%	10.6%	98.4%	11.5%	65.4%
			U-MI-FGSM		704	0.0%	90.9%	90.8%	97.3%	64.4%	80.0%	57.0%	20.2%	8.7%	97.0%	9.8%	61.6%
			UAP		303	0.0%	96.7%	95.4%	98.7%	71.3%	15.2%	12.9%	10.2%	18.2%	97.7%	42.2%	55.8%
		L_2	DF		1000	0.0%	99.6%	99.3%	99.3%	96.8%	98.9%	99.0%	68.9%	97.5%	99.3%	98.6%	95.7%
			OM		1000	0.0%	88.6%	87.7%	93.7%	70.8%	90.5%	77.9%	26.3%	4.8%	94.8%	1.0%	63.6%
	TAs	L_∞ $\epsilon=0.3$	LLC		56	0.0%	96.4%	98.2%	100.0%	73.2%	96.4%	75.0%	8.9%	23.2%	100.0%	7.1%	67.9%
			R+LLC		40	0.0%	97.5%	95.0%	97.5%	92.5%	97.5%	92.5%	25.0%	30.0%	95.0%	32.5%	75.5%
			ILLC		594	0.0%	98.7%	98.8%	99.0%	87.0%	95.1%	86.7%	25.4%	20.7%	98.5%	30.8%	74.1%
			T-MI-FGSM		864	0.0%	98.4%	97.9%	99.2%	81.9%	90.5%	72.1%	26.6%	18.8%	99.1%	22.9%	70.7%
		L_0	JSMA		764	0.0%	78.5%	74.1%	79.7%	78.5%	86.0%	73.7%	38.6%	13.7%	73.0%	35.2%	63.1%
		L_2	BLB		1000	0.0%	99.7%	99.0%	99.3%	98.7%	99.1%	99.0%	66.3%	95.2%	98.6%	98.0%	95.3%
			C[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	98.2%	99.1%	98.2%	68.3%	96.4%	98.4%	98.4%	95.5%
				[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	79.2%	90.7%	73.5%	19.8%	2.9%	81.1%	0.5%	59.1%
			E[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	98.4%	99.1%	98.5%	64.5%	96.6%	98.0%	98.3%	95.0%
				L1	1000	0.0%	99.0%	97.8%	98.3%	98.1%	99.1%	97.9%	62.0%	94.9%	96.8%	98.3%	94.2%
	Average				682.1	0.0%	90.7%	89.8%	92.6%	79.1%	85.0%	74.4%	33.8%	35.3%	91.3%	38.1%	71.0%

Average

Compare to Prior Work

An all too common paper:

CLASSIFICATION ACCURACY OF COMPLETELY DEFENSES AGAINST ADVERSARIAL ATTACKS ON MNIST

Datasets	Attack					Original Model	Defense-enhanced Models										Average
	UA/ TA	Objec- tive	Attacks		# of AEs		Adversarial Training			Gradient Masking		Input Transformation				RC	
							NAT	EAT	PAT	DD	IGR	EIT	RT	PD	TE		
MNIST	UAs	L_∞ $\epsilon=0.3$	FGSM	$\epsilon=0.3$	304	0.0%	88.2%	88.8%	94.1%	60.2%	76.6%	61.5%	26.0%	11.2%	93.8%	8.6%	60.9%
				$\epsilon=0.5$	448	0.0%	28.8%	25.9%	26.3%	29.5%	42.6%	27.9%	15.2%	2.9%	20.5%	1.6%	22.1%
			R+FGSM		342	0.0%	95.9%	95.6%	98.3%	78.4%	88.3%	77.8%	30.1%	17.3%	97.1%	19.3%	69.8%
			BIM		756	0.0%	93.0%	92.5%	97.9%	71.6%	83.6%	65.7%	21.2%	7.0%	97.6%	9.4%	63.9%
			PGD		824	0.0%	95.4%	93.8%	98.2%	74.5%	85.9%	67.7%	18.3%	10.6%	98.4%	11.5%	65.4%
			U-MI-FGSM		704	0.0%	90.9%	90.8%	97.3%	64.4%	80.0%	57.0%	20.2%	8.7%	97.0%	9.8%	61.6%
			UAP		303	0.0%	96.7%	95.4%	98.7%	71.3%	15.2%	12.9%	10.2%	18.2%	97.7%	42.2%	55.8%
		L_2	DF		1000	0.0%	99.6%	99.3%	99.3%	96.8%	98.9%	99.0%	68.9%	97.5%	99.3%	98.6%	95.7%
			OM		1000	0.0%	88.6%	87.7%	93.7%	70.8%	90.5%	77.9%	26.3%	4.8%	94.8%	1.0%	63.6%
	TAs	L_∞ $\epsilon=0.3$	LLC		56	0.0%	96.4%	98.2%	100.0%	73.2%	96.4%	75.0%	8.9%	23.2%	100.0%	7.1%	67.9%
			R+LLC		40	0.0%	97.5%	95.0%	97.5%	92.5%	97.5%	92.5%	25.0%	30.0%	95.0%	32.5%	75.5%
			ILLC		594	0.0%	98.7%	98.8%	99.0%	87.0%	95.1%	86.7%	25.4%	20.7%	98.5%	30.8%	74.1%
			T-MI-FGSM		864	0.0%	98.4%	97.9%	99.2%	81.9%	90.5%	72.1%	26.6%	18.8%	99.1%	22.9%	70.7%
		L_0	JSMA		764	0.0%	78.5%	74.1%	79.7%	78.5%	86.0%	73.7%	38.6%	13.7%	73.0%	35.2%	63.1%
		L_2	BLB		1000	0.0%	99.7%	99.0%	99.3%	98.7%	99.1%	99.0%	66.3%	95.2%	98.6%	98.0%	95.3%
			CW2	$\kappa=0$	997	0.0%	99.6%	99.0%	99.3%	98.2%	99.1%	98.2%	68.3%	96.4%	98.4%	98.4%	95.5%
				$\kappa=20$	963	0.0%	79.7%	78.7%	85.2%	79.2%	90.7%	73.5%	19.8%	2.9%	81.1%	0.5%	59.1%
			EAD	EN	1000	0.0%	99.3%	98.4%	99.0%	98.4%	99.1%	98.5%	64.5%	96.6%	98.0%	98.3%	95.0%
				L1	1000	0.0%	99.0%	97.8%	98.3%	98.1%	99.1%	97.9%	62.0%	94.9%	96.8%	98.3%	94.2%
	Average				682.1	0.0%	90.7%	89.8%	92.6%	79.1%	85.0%	74.4%	33.8%	35.3%	91.3%	38.1%	**71.0%**

Sanity-Check Conclusions

An all too common paper:

CLASSIFICATION ACCURACY OF COMPLETELY DEFENSES AGAINST ADVERSARIAL ATTACKS ON MNIST

22.1%

Datasets	Attack					Original Model	Defense-enhanced Models										Average
	UA/ TA	Objec- tive	Attacks		# of AEs		Adversarial Training			Gradient Masking		Input Transformation				RC	
							NAT	EAT	PAT	DD	IGR	EIT					
MNIST	UAs	L_∞ $\epsilon=0.3$	FGSM	$\epsilon=0.3$	304	0.0%	88.2%	88.8%	94.1%	60.2%	76.6%	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	60.9%
				$\epsilon=0.5$	448	0.0%	28.8%	25.9%	26.3%	29.5%	42.6%	27.9%	[illegible]	2.9%	[illegible]	[illegible]	22.1%
			R+FGSM		342	0.0%	95.9%	95.6%	98.3%	78.4%	88.3%	77.8%	30.1%	17.3%	97.1%	19.3%	69.8%
			BIM		756	0.0%	93.0%	92.5%	97.9%	71.6%	83.6%	65.7%	21.2%	7.0%	97.6%	9.4%	63.9%
			PGD		824	0.0%	95.4%	93.8%	98.2%	74.5%	85.9%	67.7%	18.3%	10.6%	98.4%	11.5%	65.4%
			U-MI-FGSM		704	0.0%	90.9%	90.8%	97.3%	64.4%	80.0%	57.0%	20.2%	8.7%	97.0%	9.8%	61.6%
			UAP		303	0.0%	96.7%	95.4%	98.7%	71.3%	15.2%	12.9%	10.2%	18.2%	97.7%	42.2%	55.8%
		L_2	DF		1000	0.0%	99.6%	99.3%	99.3%	96.8%	98.9%	99.0%	68.9%	97.5%	99.3%	98.6%	95.7%
			OM		1000	0.0%	88.6%	87.7%	93.7%	70.8%	90.5%	77.9%	26.3%	4.8%	94.8%	1.0%	63.6%
	TAs	L_∞ $\epsilon=0.3$	LLC		56	0.0%	96.4%	98.2%	100.0%	73.2%	96.4%	75.0%	8.9%	23.2%	100.0%	7.1%	67.9%
			R+LLC		40	0.0%	97.5%	95.0%	97.5%	92.5%	97.5%	92.5%	25.0%	30.0%	95.0%	32.5%	75.5%
			ILLC		594	0.0%	98.7%	98.8%	99.0%	87.0%	95.1%	86.7%	25.4%	20.7%	98.5%	30.8%	74.1%
			T-MI-FGSM		864	0.0%	98.4%	97.9%	99.2%	81.9%	90.5%	72.1%	26.6%	18.8%	99.1%	22.9%	70.7%
		L_0	JSMA		764	0.0%	78.5%	74.1%	79.7%	78.5%	86.0%	73.7%	38.6%	13.7%	73.0%	35.2%	63.1%
		L_2	BLB		1000	0.0%	99.7%	99.0%	99.3%	98.7%	99.1%	99.0%	66.3%	95.2%	98.6%	98.0%	95.3%
			CW2	$\kappa=0$	997	0.0%	99.6%	99.0%	99.3%	98.2%	99.1%	98.2%	68.3%	96.4%	98.4%	98.4%	95.5%
				$\kappa=20$	963	0.0%	79.7%	78.7%	85.2%	79.2%	90.7%	73.5%	19.8%	2.9%	81.1%	0.5%	59.1%
			EAD	EN	1000	0.0%	99.3%	98.4%	99.0%	98.4%	99.1%	98.5%	64.5%	96.6%	98.0%	98.3%	95.0%
				L1	1000	0.0%	99.0%	97.8%	98.3%	98.1%	99.1%	97.9%	62.0%	94.9%	96.8%	98.3%	94.2%
	Average				682.1	0.0%	90.7%	89.8%	92.6%	79.1%	85.0%	74.4%	33.8%	35.3%	91.3%	38.1%	71.0%

An all too common paper:

CLASSIFICATION ACCURACY OF COMPLETELY DEFENSES AGAINST ADVERSARIAL ATTACKS ON MNIST

Datasets	Attack					Original Model	Defense-enhanced Models										Average
	UA/TA	Objective	Attacks		# of AEs		Adversarial Training			Gradient Masking		Input Transformation				RC	
							NAT	EAT	PAT	DD	IGR	EIT	RT	PD	TE		
MNIST	UAs	L_∞ $\epsilon=0.3$	FGSM	$\epsilon=0.3$	304	0.0%	88.2%	88.8%	94.1%	60.2%	76.6%	61.5%	26.0%	11.2%	93.8%	8.6%	60.9%
				$\epsilon=0.5$	448	0.0%	28.8%	25.9%	26.3%	29.5%	42.6%	27.9%	15.2%	2.9%	20.5%	1.6%	22.1%
			R+FGSM		342	0.0%	95.9%	95.6%	98.3%	78.4%	88.3%	77.8%	30.1%	17.3%	97.1%	19.3%	69.8%
			BIM		756	0.0%	93.0%	92.5%	97.9%	71.6%	83.6%	65.7%	21.2%	7.0%	97.6%	9.4%	63.9%
			PGD		824	0.0%	95.4%	93.8%	98.2%	74.5%	85.9%	67.7%	18.3%	10.6%	98.4%	11.5%	65.4%
			U-MI-FGSM		704	0.0%	90.9%	90.8%	97.3%	64.4%	80.0%	57.0%	20.2%	8.7%	97.0%	9.8%	61.6%
			UAP		303	0.0%	96.7%	95.4%	98.7%	71.3%	15.2%	12.9%	10.2%	18.2%	97.7%	42.2%	55.8%
		L_2	DF		1000	0.0%	99.6%	99.3%	99.3%	96.8%	98.9%	99.0%	68.9%	97.5%	99.3%	98.6%	95.7%
			OM		1000	0.0%	88.6%	87.7%	93.7%	70.8%	90.5%	77.9%	26.3%	4.8%	94.8%	1.0%	63.6%
	TAs	L_∞ $\epsilon=0.3$	LLC		56	0.0%	96.4%	98.2%	100.0%	73.2%	96.4%	75.0%	8.9%	23.2%	100.0%	7.1%	67.9%
			R+LLC		40	0.0%	97.5%	95.0%	97.5%	92.5%	97.5%	92.5%	25.0%	30.0%	95.0%	32.5%	75.5%
			ILLC		594	0.0%	98.7%	98.8%	99.0%	87.0%	95.1%	86.7%	25.4%	20.7%	98.5%	30.8%	74.1%
			T-MI-FGSM		864	0.0%	98.4%	97.9%	99.2%	81.9%	90.5%	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	70.7%
		L_0	JSMA		764	0.0%	78.5%	74.1%	79.7%	78.5%	86.0%	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	63.1%
		L_2	BLB		1000	0.0%	99.7%	99.0%	99.3%	98.7%	99.1%	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	95.3%
			CW2	$\kappa=0$	997	0.0%	99.6%	99.0%	99.3%	98.2%	99.1%	98.2%	[illegible]	[illegible]	[illegible]	98.4%	95.5%
				$\kappa=20$	963	0.0%	79.7%	78.7%	85.2%	79.2%	90.7%	73.5%	19.8%	2.9%	81.1%	0.5%	59.1%
			EAD	EN	1000	0.0%	99.3%	98.4%	99.0%	98.4%	99.1%	98.5%	64.5%	96.6%	98.0%	98.3%	95.0%
				L1	1000	0.0%	99.0%	97.8%	98.3%	98.1%	99.1%	97.9%	62.0%	94.9%	96.8%	98.3%	94.2%
	Average				682.1	0.0%	90.7%	89.8%	92.6%	79.1%	85.0%	74.4%	33.8%	35.3%	91.3%	38.1%	**71.0%**

95.5%

Making errors in
defense evaluations is *okay*.

Making errors in
attack evaluations is not.

Breaking a defense
is useful ...

... teaching a lesson
is better

Exciting new directions

Exciting new directions

Decision-Based Adversarial Attacks: Reliable Attacks Against Black-Box Machine Learning Models

Wieland Brendel*, Jonas Rauber* & Matthias Bethge
Werner Reichardt Centre for Integrative Neuroscience,
Eberhard Karls University Tübingen, Germany
{wieland,jonas,matthias}@bethgelab.org

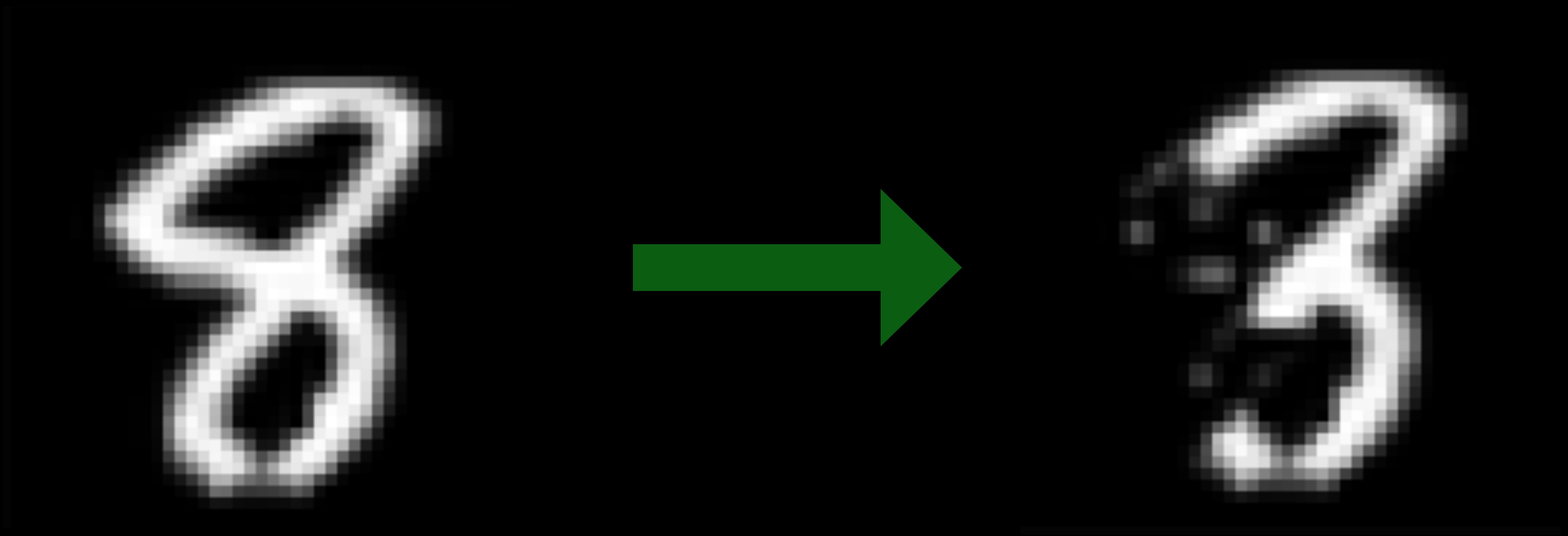
Exciting new directions

EXCESSIVE INVARIANCE CAUSES ADVERSARIAL VULNERABILITY

Jörn-Henrik Jacobsen [1*], Jens Behrmann[1,2], Richard Zemel[1], Matthias Bethge[3]
[1]Vector Institute and University of Toronto
[2]University of Bremen, Center for Industrial Mathematics
[3]University of Tübingen

Exciting new directions

Exciting new directions

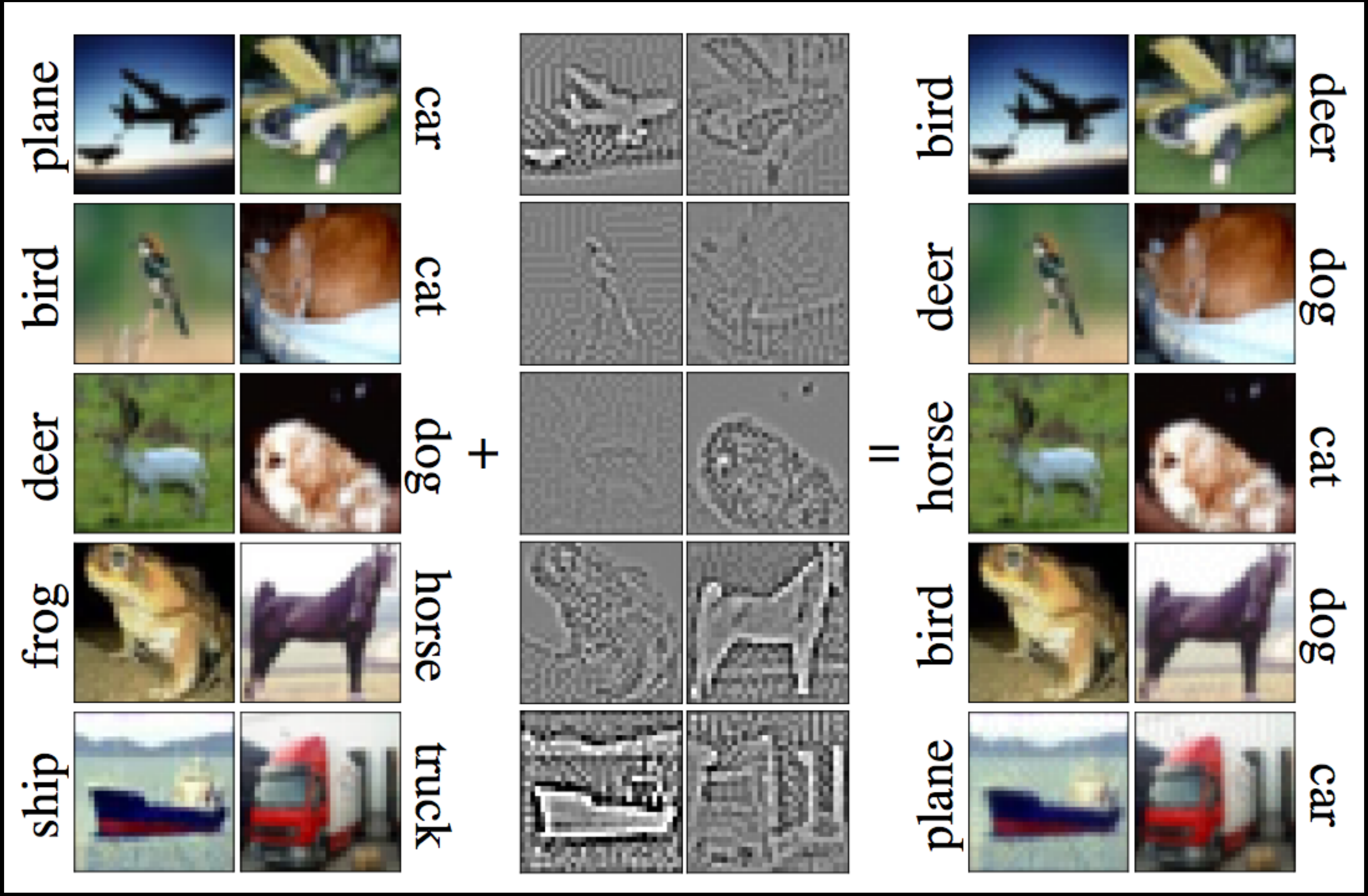


Exciting new directions

Wasserstein Adversarial Examples via Projected Sinkhorn Iterations

Eric Wong[1] **Frank R. Schmidt**[2] **J. Zico Kolter**[3 4]

Exciting new directions



Act VI
Conclusions

Research new topics

Do good science

Progress is learning

Questions?

nicholas@carlini.com https://nicholas.carlini.com

References

Biggio et al. Evasion Attacks on Machine Learning at Test Time.
https://arxiv.org/abs/1708.06131
Jaconbsen et al. Exploiting Excessive Invariance caused by Norm-Bounded Adversarial Robustness
https://arxiv.org/abs/1903.10484
Carlini et al. On Evaluating Adversarial Robustness.
https://arxiv.org/abs/1902.06705
Chou et al. SentiNet: Detecting Physical Attacks Against Deep Learning Systems.
https://arxiv.org/abs/1812.00292
Shumailov et al. Sitatapatra: Blocking the Transfer of Adversarial Samples.
https://arxiv.org/abs/1901.08121
Ilyas et al. Adversarial Examples Are Not Bugs, They Are Features.
https://arxiv.org/abs/1905.02175
Brendel et al. Decision-Based Adversarial Attacks: Reliable Attacks Against Black-Box Machine Learning
https://arxiv.org/abs/1712.04248
Wong et al. Wasserstein Adversarial Examples via Projected Sinkhorn Iterations
https://arxiv.org/abs/1902.07906.