Deep Learning: (still) Not Robust

Nicholas Carlini Google

Better Language Models and Thei Implications

We've trained a large-scale unsupervise model which generates coherent paragr text, achieves state-of-the-art performa many language modeling benchmarks, a performs rudimentary reading comprehe machine translation, question answering summarization—all without task-specifi

February 14, 2019 24 minute read

Deep Speech 2: End-to-l English an

Baidu Research -Dario Amodei, Rishita Anubhai, Eric Batten Jingdong Chen, Mike Chrzanowski, Adam Linxi Fan, Christopher Fougner, Tony Har Libby Lin, Sharan Narang, Andrew Ng, S Sanjeev Satheesh, David Seetapun, Shubho S Bo Xiao, Dani Yogatan

> We show that an end-to-end deep lea either English or Mandarin Chinese sp cause it replaces entire pipelines of han works, end-to-end learning allows us to ing noisy environments, accents and different languages. Key to our approach is our application of HPC techniques, resulting in a 7x speedup over our previous system [26]. Because of this efficiency, experiments that previously took weeks now run in days. This enables us to iterate more quickly to identify superior architectures and algorithms. As a result, in several cases, our system is competitive with the transcription of human workers when benchmarked on standard datasets. Finally, using a technique called Batch Dispatch with GPUs in the data center, we show that our system can be inexpensively deployed in an online setting, delivering low latency when serving users at scale.

Facebook

Introducing the First AI Model That **Translates 100 Languages Without Relying on English**

October 19, 2020 By Angela Fan, Research Assistant

Ab













adversarial perturbation

88% tabby cat



99% guacamole





football Al Camera Ruins Soccer Game For Fans After Mistaking **Referee's Bald Head For Ball**



Share on Facebook

Share on Twitter 1







Adversaria Distribution Shifts

र्<u>ति जनानान</u>न

💋 💿 👟 🌮 🎒 🙁 🎯 🔼 🔇 🖉

AI Camera Ruins Soccer Game For Fans After Mistaking **Referee's Bald Head For Ball**





99% guacamole





Natural 19 Distribution Shifts

Acversaria Distribution Shifts

Distribution Shifts



Acversaria Distribution Shifts



NeurIPS'20, with Florian Tramer, Wieland Brendel, Aleksander Madry

Acversaria Distribution Shifts



Adversarial (n.) Defn: "involving or characterized by conflict or opposition."

GIVEN

FIND

SUCH THAT

a neural network f an input to the network x

a new input x'

f(x') is classified incorrectly x and x' are close



Adversarial Accuracy

Probability an adversary can succeed at this game



On Adaptive Attacks to Adversarial Example Defenses

Florian Tramèr^{*} Stanford University

- 5 k-Winners Take All
- 6 The Odds are Odd
- 7 Are Generative Classifiers More Robust?
- 8 Robust Sparse Fourier Transform
- 9 Rethinking Softmax Cross Entropy
- 10 Error Correcting Codes
- 11 Ensemble Diversity
- 12 EMPIR
- 13 Temporal Dependency
- 14 Mixup Inference
- 15 ME-Net
- 16 Asymmetrical Adversarial Training
- 17 Turning a Weakness into a Strength
- 18 Conclusion

Nicholas Carlini^{*} Google Brain Aleksander Mądry MIT

> 8 11 $\mathbf{14}$ 17 $\mathbf{18}$ $\mathbf{20}$ $\mathbf{22}$ $\mathbf{24}$ $\mathbf{25}$ $\mathbf{28}$ $\mathbf{30}$ $\mathbf{32}$ 3538

Wieland Brendel^{*}

University of Tübingen

We evaluated 13 defenses proposed at (ICLR|ICML|NeurIPS) 20(18|19|20)

All were broken. Adversarial accuracy of roughly 0%.

Random Direction

Random Direction



<u>Random</u> Direction

Random Direction









Bandom Direction

Random Direction





Bandom Direction

Adversarial Direction



What do defenses do?







Our paper: Adaptive Attacks





I'm not going to tell you how we broke them.

... it's quite boring.







Instead let's talk about the context of this paper

Previously



New Idea 1 -New Idea 2 -New Idea 3

New Idea A New Idea B New Idea C

Defensive Distillation is Not Robust to Adversarial Examples

We identi

ent maskii

sense of s

examples.

gradients

based atta

effect can

teristic be

and for ea

dients we

to overco

certified w

we find ob

rence, wit

gradients.

vent 6 coi

threat mo

1. Introduct

In response to t

sarial example

there has been

fenses to increa

progress has b

against adversa

the adversary

solution has no

As benchmark

tacks (e.g., Ki

Adversarial Examples Are Not Easily Detected: Bypassing Ten Detection Methods

Abst

We sh

more

unpro

It is a

they

sive d

to mal

versar

tillatio

ificati

ample

strate

advers

inputs

attack

magni

two of

achiev

of im:

2

2.1

We as

sarial

briefly

param

last la

layer

 $F(\theta, x)$

cation

proba

 $C(\theta, x)$

tion o

we ar

greyse

Let

Dis

In t

1

ABSTRAC

Neural netwo ples: inputs the In order to bet survey ten rec pare their effic new loss funct nificantly har properties bel fact not. Finally

Recent years h Neural netwo driving force b been demonstr [38], to beatin cars [6]. sification. Wh learning appro who attempts x, an adversar has a different adversarial ex nearly all dom proposing mai ples correctly these defenses correctly. Due to this detect them in seven papers pare their effi With new atta evaded by an a

datasets, the a Permission to mal classroom use is g for profit or comm on the first page. author(s) must be republish, to post of and/or a fee. Requ AISec'17, Novembe © 2017 Copyright tion for Computin ACM ISBN 978-1-

https://doi.org/10.

future propose

1 INTRO

In this pape The researc

Abstract MagNet and

posed as a de we can cons defenses with

1 Introd

It is an oper they will be cently, three networks rol

> MagNe neural n through ial exar to lie of rectly, v the data classific the whi MagNe adversa the para rameter

 An efficiency neural r amples during function but clair

 Adversa Carlini & Wag

MagNet and "Efficient Defenses Against Adversaria are Not Robust to Adversarial Examples

Obfuscated Gradients Give a False Ser Circumventing Defenses to Adversar

On the Robustness of the CVPR 2018 V

Robust

Abstract-No.

I. ATTACKING "ATTACKS MEET INTE

Is AmI (A

AmI (Attacks meet Interpretability) is an defense [3] to detect [1] adversarial exa recognition models. By applying interprito a pre-trained neural network, AmI ide neurons. It then creates a second augmer with the same parameters but increases the of important neurons. AmI rejects inputs and augmented neural network disagree.

We find that this defense (presented at a a spotlight paper-the top 3% of submiss ineffective, and even defense-oblivious¹ detection rate to 0% on untargeted attacks. more robust to untargeted attacks than the network. Figure 1 contains examples of a that fool the AmI defense. We are incred authors for releasing their source code² w We hope that future work will continue to by publication time to accelerate progress

	T
	Va
	Sat
2	

S1.

Neural netv adversarial two white-2018 and fir existing tec of the defer

1. Introducti

sarial examples (Two defenses that this problem: "I Deflection" (Pral versarial Attacks Denoiser" (Liao

In this note, we in the white-boy examples that re ImageNet datas a small ℓ_{∞} pert considered in the A. Evaluation

Training neural

Comment on Biologically inspired protection of deep networks from adversarial attacks

ON THE LIMITATION OF LOCAL INTRINSIC DIMEN-SIONALITY FOR CUADACTERIZING THE SUDOBACEO OF

 $^{1}Wern\epsilon$

А Adversarial Risk and the Dangers of Evaluating Against Weak Attacks Τź

The Efficacy of SHIELD under Different Threat Models

Paper Type: Appraisal Paper of Existing Method

Cory Cornelius cory.cornelius@intel.com

Nilaksh Das nilakshdas@gatech.edu

Shang-Tse Chen schen351@gatech.edu

Evaluating and Understanding the Robustness of **Adversarial Logit Pairing**

Andrew Ilyas* Logan Engstrom* Anish Athalye* Massachusetts Institute of Technology {engstrom, ailyas, aathalye}@mit.edu

Abstract

We evaluate the robustness of Adversarial Logit Pairing, a recently proposed defense against adversarial examples. We find that a network trained with Adversarial Logit Pairing achieves 0.6% correct classification rate under targeted adversarial attack, the threat model in which the defense is considered. We provide a brief overview of the defense and the threat models/claims considered, as well as a discussion of the methodology and results of our attack. Our results offer insights into the reasons underlying the vulnerability of ALP to adversarial attack, and are of general interest in evaluating and understanding adversarial defenses.

1 Contributions

For summary, the contributions of this note are as follows:

 Robustness: Under the white-box targeted attack threat model specified in Kannan et al., we upper bound the correct classification rate of the defense to 0.6% (Table 1). We also perform targeted and untargeted attacks and show that the attacker can reach success rates of 98.6% and 99.9% respectively (Figures 1, 2).

A rece gradient-l highly sat the attack stabilisati has yet to limitation Evaluat between ne robust beca for gradien in cases wi attacks tha of the grad A recen against gra likely a side stable com of the grad In a fir Perceptron layer activa sigmoid, ze network for the MLP w In r attack with exa verified the of a effc the In high and exactly zer exa directly wi und elements of gen Mo raining sic gen milla turatedTable 1: A naive appli whe FGSM bas dra (20)

DN

ple mo ach the me to mo the ABSTRA as In this appr and compression and ial attacks o stra by niq dec def lati mo . Intr Deep lea ing and ing prob speech game pl able suc properti instead of be Researc tions to tremely 2017) F percepti

Th

pro

at KDD 201 studied in t adversary is pre-process used in the threat and e

degree of in full white-bo original wo an adaptive of the Proje gradient-bas learning res ensemble fro pre-trained targeted PG Shield ense 48.9% if the

ensemble w tion in the c scratch are l gray-box sc

ACM Refere











New Idea 1 -New Idea 2 -New Idea 3 New Idea 95

New Idea A New Idea B New Idea C

just reuse one



On Adaptive Attacks to Adversarial Example Defenses

Florian Tramèr^{*} Stanford University Nicholas Carlini^{*} Google Brain

Wieland Brendel^{*} University of Tübingen

Aleksander Mądry MIT

- 5 k-Winners Take All
- The Odds are Odd 6
- Are Generative Classifiers More Robust? 7
- Robust Sparse Fourier Transform 8
- Rethinking Softmax Cross Entropy 9
- 10 Error Correcting Codes
- 11 Ensemble Diversity
- 12 EMPIR
- 13 Temporal Dependency
- 14 Mixup Inference
- 15 ME-Net
- 16 Asymmetrical Adversarial Training
- 17 Turning a Weakness into a Strength
- 18 Conclusion

Another weakness of the paper is that defenses are broken by existing techniques. Indeed, at the end of the analysis, most of the defenses are broken either by using EOT, BPDA, or by tuning the parameters of existing attacks such as PGD. All this techniques already exist in the literature [1,2,3,4]; hence the technical part is not novel.

8

11

14

17

18

 $\mathbf{20}$

 $\mathbf{22}$

 $\mathbf{24}$

 $\mathbf{25}$

 $\mathbf{28}$

30

 $\mathbf{32}$

35

38




1. Code is now always available

2. Adaptive attacks are at least attempted

The problem is methodological





for example ... one paper's attack

$$\mathcal{L}_1 = \underbrace{\mathcal{L}(h(\mathbf{x}'), \mathbf{p}^{\mathrm{adv}})}_{\mathrm{misclassify} \, \mathbf{x}' \, \mathrm{as} \, y_t},$$







$$)-h(\mathbf{x}'+\epsilon)\|_1],$$

 $\mathcal{L}_3 = \mathbb{E}_{y' \sim \text{Uniform}, y' \neq y_t} [\mathcal{L}(h(\mathbf{x}' - \alpha \delta_{y'}), y')],$

$$\mathcal{L}_4 = -\mathcal{L}(h(\mathbf{x}' + \alpha \delta_{y_t}), y_t).$$

$$\mathcal{L}_2 + \mathcal{L}_3 + \mathcal{L}_4.$$

for example ... one paper's attack

$$\mathcal{L}_1 = \underbrace{\mathcal{L}(h(\mathbf{x}'), \mathbf{p}^{\mathrm{adv}})}_{\mathrm{misclassify} \, \mathbf{x}' \, \mathrm{as} \, y_t},$$







$$)-h(\mathbf{x}'+\epsilon)\|_1],$$

 $\mathcal{L}_3 = \mathbb{E}_{y' \sim \text{Uniform}, y' \neq y_t} [\mathcal{L}(h(\mathbf{x}' - \alpha \delta_{y'}), y')],$

$$\mathcal{L}_4 = -\mathcal{L}(h(\mathbf{x}' + \alpha \delta_{y_t}), y_t).$$

$$\mathcal{L}_2 + \mathcal{L}_3 + \mathcal{L}_4.$$

for example ... our attack

$\mathcal{L}_1 = \mathcal{L}(h(\mathbf{x}'), \mathbf{p}^{\mathrm{adv}})$

misclassify \mathbf{x}' as y_t



Acversaria Distribution Shifts

Acversaria Distribution Shifts

Distribution Shifts

Natural

Distribution Shifts

Rohan Taori, Achal Dave, Vaishaal Shankar, Benjamin Recht, Ludwig Schmidt



Natural (adj.) Defn: "existing in or caused by nature"

What we want

1. Someone wants to know what breed of dog they just saw on the street 2. They take out their phone 3. Open up the camera app

- 4. Take a picture, and run a ResNet on the image

What we have

1. Someone wants to know what breed of dog they just saw on the street 2. They take out their phone 3. Open up the camera app 4. Close the camera app. Open up the browser. Visit <u>http://image-net.org/</u>. Download the ILSVRC2012 test set. Select an image of a dog uniformly at random. Ask the resnet model to classify that random image. Ignore the real dog.



Constructing "natural" datasets

Benjamin UC Berl

ObjectNet: A large-scale bias-co pushing the limits of object re-

We buil the focus of re-used tes extent curre and find ac accuracy ga suggest tha generalize t

Andrei Barbu* MIT, CSAIL & CBMM

David Mayo* MIT, CSAIL & CBMM

Christopher Wang MIT, CSAIL

Dan Gutfreund MIT-IBM Watson AI

Joshua Ten MIT, BCS &

We study the robustness of image part of this study, we construct two containing a total of 57,897 images Our datasets were derived from Im: re-annotated by human experts for i pre-trained on ImageNet and show a datasets. Additionally, we evaluate the induce both classification as well as lo of 14 points. Our analysis demonstra substantial and realistic challenge to that require both reliable and low-la

Abstract

We collect a large real-world test set, ObjectNet, for ob where object backgrounds, rotations, and imaging vi scientific experiments have controls, confounds whic to ensure that subjects cannot perform a task by exp the data. Historically, large machine learning and co lacked such controls. This has resulted in models that datasets and perform better on datasets than in real tested on ObjectNet, object detectors show a 40-459 respect to their performance on other benchmarks, d Controls make ObjectNet robust to fine-tuning show

increases. We develop a highly automated platform that enables gathering datasets with controls by crowdsourcing image capturing and annotation. ObjectNet is the same size as the ImageNet test set (50,000 images), and by design does not come paired with a training set in order to encourage generalization. The dataset is both easier than ImageNet - objects are largely centered and unoccluded - and harder, due to the controls. Although we focus on object recognition here, data with controls can be gathered at scale using automated tools throughout machine learning to generate datasets that exercise models in new ways thus providing valuable feedback to researchers. This work opens up new avenues for research in generalizable, robust, and more human-like computer vision and in creating datasets where results are predictive of real-world performance.

Do Image Classifiers Generalize Across Time?

Deva Ramanan CMU

Vaishaal Shankar^{*} UC Berkeley

Achal Dave^{*}

Rebecca Roelofs

Natural Adversarial Examples

Dan Hendrycks UC Berkeley hendrycks@berkeley.edu

Kevin Zhao* University of Washington kwzhao@cs.washington.edu

Steven Basart' University of Chicago steven@ttic.edu

Jacob Steinhardt UC Berkeley jsteinhardt@berkeley.edu

Dawn Song UC Berkeley dawnsong@berkeley.edu

Abstract

We introduce natural adversarial examples-real-world, unmodified, and naturally occurring examples that cause machine learning model performance to substantially degrade. We introduce two new datasets of natural adversarial examples. The first dataset contains 7,500 natural adversarial examples for ImageNet classifiers and serves as a hard ImageNet classier test set called IMAGENET-A. We also curate an adversarial out-of-distribution detection dataset called IMAGENET-O, which to our knowledge is the first out-of-distribution detection dataset created for ImageNet models. These two datasets provide new ways to measure model robustness and uncertainty. Like ℓ_p adversarial examples, our natural adversarial examples transfer to unseen black-box models. For example, on IMAGENET-A a DenseNet-121 obtains around 2% accuracy, an accuracy drop of approximately 90%, and its out-of-distribution detection performance on IMAGENET-O is near random chance levels. Popular training techniques for improving robustness have little effect, but some architectural changes provide mild improvements. Future research is required to enable generalization to natural adversarial examples.



Figure 1: Natural adversarial examples from IMAGENET-A and IMAGENET-O. The black text is the actual class, and the red text is a ResNet-50 prediction and its confidence.



Do ImageNet Classifiers Generalize to ImageNet?

Benjamin Recht^{*} UC Berkeley Rebecca Roelofs UC Berkeley Ludwig Schmidt UC Berkeley Vaishaal Shankar UC Berkeley

Abstract

We build new test sets for the CIFAR-10 and ImageNet datasets. Both benchmarks have been the focus of intense research for almost a decade, raising the danger of overfitting to excessively re-used test sets. By closely following the original dataset creation processes, we test to what extent current classification models generalize to new data. We evaluate a broad range of models and find accuracy drops of 3% - 15% on CIFAR-10 and 11% - 14% on ImageNet. However, accuracy gains on the original test sets translate to larger gains on the new test sets. Our results suggest that the accuracy drops are not caused by adaptivity, but by the models' inability to generalize to slightly "harder" images than those found in the original test sets.

Do ImageNet Classifiers Generalize to ImageNet?

Benjamin Recht^{*} Rebecca Roelofs Ludwig Schmidt UC Berkeley UC Berkeley UC Berkeley

Abstract

We build new test sets for the CIFAR-10 and ImageNet datasets. Both benchmarks have been the focus of intense research for almost a decade, raising the danger of overfitting to excessively re-used test sets. By closely following the original dataset creation processes, we test to what extent current classification models generalize to new data. We evaluate a broad range of models and find accuracy drops of 3% - 15% on CIFAR-10 and 11% - 14% on ImageNet. However, accuracy gains on the original test sets translate to larger gains on the new test sets. Our results suggest that the accuracy drops are not caused by adaptivity, but by the models' inability to generalize to slightly "harder" images than those found in the original test sets.

Vaishaal Shankar UC Berkeley

This research study is being conducted by Ben Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar from UC Berkeley. For question about this study, please contact ludwig@berkeley.edu and roelofs@cs.berkeley.edu. In this study, we will ask you to indicate whether given image belong to a certain object category. Occasionally, the images may contain disturbing or adult content. We would like to remind you that participation in our study is voluntary and that you can withdraw from the study at any time.

Which of these images contain at least one object of type

Definition: an English breed slightly larger than the American foxhounds originally used to hunt in packs

Task:

For each of the following images, check the box next to an image if it contains at least one object of type English foxhound. Select an image if it contains the object regardless of occlusions, other objects, and clutter or text in the scene. Only select images that are photographs (no drawings or paintings).

Please make accurate selections!

If you are unsure about the object meaning, please also consult the following Wikipedia page(s): https://en.wikipedia.org/wiki/English_Foxhound

If it is impossible to complete a HIT due to missing data or other problems, please return the HIT.













English foxhound









Now we have a new dataset. Identical in every way to the original. How do models do on this new dataset?





ImageNet









ImageNet

New test accuracy (top-1, %) 40 60 70 Original test accuracy (top-1, %)



80

Adaptive Overfitting







ImageNet

Are there any ways to increase robustness to this distribution shift?



45	2.5	12	7.3	50	9.8	12	58	64	38	27	36	50	29	33	22	8.3	41	31	17	58	19	26	41	24	36	58	28	23	35	36 3	27 28	21	18	15	41 :	27 2	29	52	19	22	40 :	26 25	8 6	0 69	74	
						10							~		-		10		-		10															~										
40	25	11	1.4	49	9.6	12	29	64	36	21	30	50	29	33	20	8.3	42	31	20	57	18	2/	43	24	30	56	28	24	30	30 3	21 28	21	18	10	42 2	21 3	33 :	92	19	23 4	*0 :	ko 29		9 08	/4	
44	23	11	7.4	49	9.5	12	59	64	38	26	36	50	30	33	20	9.7	40	31	19	57	18	27	42	24	36	58	28	23	35	36 2	27 28	20	19	16	40 2	27 3	32 8	52	18 :	23 4	40 2	25 29) 59	9 69	74	
45	1.7	9.8	7.7	43	9	20	60	65	73	37	43	52	40	42	31	27	55	37	50	58	28	30	45	34	38	61	57	33	42	42 :	36 36	28	26	19	50 :	34 8	50 !	53 :	24	29	43	31 33	58	9 68	74	
42	1.8	8.9	7.1	46	8.1	17	59	63	37	29	41	44	27	35	21	15	49	35	27	57	19	25	46	25	35	58	28	28	40	33 :	26 32	23	21	18	48 :	30 3	36 !	53 :	21	23	43 :	28 25) 58	9 68	74	
46	2.3	10	8.2	41	9.4	19	58	63	59			74	40	41	24	20	47	35	Λ	53	23	32	44	28	36	58	44	30	38	63	37 33	22	23	15	44 :	31 3	39 4	48 2	20	30	42 2	26 30) 50	9 68	74	
44	1.9	10	7	54	9	17	59	65	42	32	41	8	32	38	28	и	48	35	44	58		28	47		30	60		3	39	37	50	27	22	20	10		4 4	54 3	24	25	43 :	32 3°	1 50	9 68	73	
43	2.6	12	7.1	52	9.6	12	58	63	37	26	36	7	27	32	20	10	41	34				25		3	36	58	27	3	35	34	20	,	18	15	40 2	2 3	31 4	51	18	23	40 :	24 28	3 58	9 68	74	
43	2.1	7.7	6.2	40	6.9	11	60	64	39	6.		47		53	17	4.3	40	.3	22	57			42	2		58	28	24		34	2	20		14	41 2	2	э. е	51	18 :	22	40 :	25 25	9 8	9 67	74	
44	1.6	9.7	7.2	53	9.1	17	59	64	43	33	40	47	32	39	27	22	47	36	41	58	20	29	46	-33	38	60	33	30	39	37 ;	30 34	26	22	21	46 :	32 4	44 ÷	54 2	24	26	42 3	32 3 [.]	1 5	9 68	74	
45	1.7	9.7	7.2	55	9.2	16	58	63	41	33	39	47	32	38	27	22	49	37	42	58	25	29	46	33	39	59	31	29	39	36 3	31 33	27	22	22	46 :	32 4	43 (54 2	25	26	43 :	32 3	3 5	9 68	74	
47	1.3	13	8.9	54	12	16	56	62	37	32	41	43	32	39	28	26	51	35	41	56	27	23	43	33	36	58	29	30	40	33 3	28 33	23	24	17	47 :	31 4	41 4	51 :	21	23	41 :	28 3	0 5	8 67	73	
43	1.7	8.9						62	3		36	45					43	32	30	56	16						28	25	34	34		18	19						17 :	22		3	0 5	7 66	73	
52	2	70		70	70	20		61	2		51			43	37	3		38	62	57	37			46	42			32	49	11		34	38	31	62		50 <u>(</u>	53	33	29			5	8 67	73	
41	- 10	44		50		15		=0	-		25			30	47	75	50	~	40					20				24	99				45	10			28	40	15					7 66		
	1.0			50	8.4				-				2	30		7.0	30	20	10		14			20	34 00			~	20				0	12				40 50	~						12	
40	1.7	14						60	Ð		41		30	30				33	54	63	33			37	32	01		20	39		33 25		22	22	47 47		o1 :	58 7	~			13		5 67	73	
41	1.1	10		52	9.1	17		62	3		41		Ģ	38	25	18		34	41	56	23			30	36	E.		28	4/				23	18	46		42 4	53 ;	21					67	73	
37	2.3	9.6		38	9.7	17		60	3		44			35	14			33	30	54	13			19	34			26			26 30			12	39	3	36 (50	7		38 3	23 20		65	71	
41	1.6	14					-35	59	з		37	39				4		28	21	52	15			_			23	21		29 2	23 25	17		11	38	2	29 4	47			36 2	22 27		75	71	
49	2.8	11	8.1	46	10	14	50	58	38	45	51	41	32	48	23	21	53	38	72	53	23	21	41	30	40	62	34	41	51	38 :	34 44	24	32	19	50 :	35 (68 (55 :	23	27	46 :	31 38	57	7 66	72	
46		69	66	26	4.8	28	57	65	19	30	56	17	49	34	39	27	65	39	58	61	38	51	56	45	42	60	10	23	49	5.8	42 26	38	32	29	59 :	31 5	55 (53 :	36 4	42 -	48 4	13 34	52	5 65	71	
48	2.4	68	48	68	68	19	50	59	18	33	50	17	39	40	43	41	65	38	61	53	42	28	51	51	42	56	11	30	48	7.8	35 33	41	37	39	61 :	35 5	59 4	49 4	40	28	4 5 4	19 37	· 58	5 64	71	
37	1.9	10	8.1	42	9	13	52	57	27	22	39	34	21	27	14	9.5	41	30	24	50	13	18	39	20	31	52	22	22	37	27 :	21 26	15	18	11	40 2	27 3	31 4	46	14	18	37 2	20 25	5 5	4 63	70	
46	2.5	9.3	7.8	35	8.9	16	48	58	37	49	49	42	25	48	21	17	46	72	46	52	19	25	40	25	53	60	31	37	51	37 :	30 39	21	29	15	47 (63 4	47 (52	19	29	44 :	25 4(3 50	6 65	71	
43		65	63	26	3.2	27	54	62	18	28	53	16	46	31	32	22	62	36	55	58	32	49	54	39	39	57	8.5	20	46	4.8	38 24	35	29	26	56 :	29 (52 (50 :	34 4	40	46 4	41 3 [.]	5	3 62	69	
47	22	64	53	64	64	18	46	54	16	31	50	15	37	38	46	45	63	37	59	49	46	29	50	54	41	52	7.7	28	47	5	33 31	45	36	43	60 :	34 8	56 4	45 4	45	28	45	52 36	5 5	1 60	68	
41	2.3	11	6.1	47	8.5	15	67	51	30	18	25	35	27	26	31	28	46	23	33	54	29	18	35	33	23	43	20	15	24	23 2	21 19	21	11	16	41 1	18 3	30 4	49	19	18	32 2	23 1:	5 5	1 60	66	

Formalization:

Effective Robustness










So what helps?





Possible explanations

1. It's just a harder dataset 2. Adaptive overfitting 3. Distribution shift 4. It's just a weird dataset







Possible explanations

1. It's just a harder dataset 2. Adaptive overfitting 3. Distribution shift 4. It's just a weird dataset



Does synthetic

natural robustness

Gaussian Noise Shot Noise Impulse Noise Defocus Blur Frosted Glass Blur



Motion Blur

Zoom Blur



Brightness

Contrast



Figure 1: Our IMAGENET-C dataset consists of 15 types of algorithmically generated corruptions from noise, blur, weather, and digital categories. Each type of corruption has five levels of severity, resulting in 75 distinct corruptions. See different severity levels in Appendix **B**.

Snow

Frost

Fog

Elastic

Pixelate

JPEG



If you wan to increase robustness, you can ...

1. Train on more data 2. Train on the distribution shift you care about

If you wan to increase robustness, you can...

1. Train on more data 2. Train on the distribution shift you care about 3. Train on the distribution shift you care about

If you wan to increase robustness, you can...

1. Train on more data 2. Train on the distribution shift you care about 3. Train on the distribution shift you care about 4. Train on the distribution shift you care about

And this is what makes adversarial/natural shifts hard to solve



Neural networks are (still) not robust

nicholas@carlini.com

https://nicholas.carlini.com

