What causes adversarial examples?

Edgar Dobriban
Wharton, UPenn

November 29, 2020

Overview

Overview

Theoretical explanations proposed

Overview

Overview

Theoretical explanations proposed

Robustness

- Want our results to not be affected by accidental perturbations of the data
- Long history: median (middle ages), L_1 regression (Boscovich, Laplace), M-estimation (Huber, Rousseeuw, ...), ...
- This talk: test-time (adversarial) robustness of prediction methods (?2004–2013–...)

Robustness

- Want our results to not be affected by accidental perturbations of the data
- Long history: median (middle ages), L_1 regression (Boscovich, Laplace), M-estimation (Huber, Rousseeuw, ...), ...
- This talk: test-time (adversarial) robustness of prediction methods (?2004–2013–...)

Robustness

- Want our results to not be affected by accidental perturbations of the data
- Long history: median (middle ages), L_1 regression (Boscovich, Laplace), M-estimation (Huber, Rousseeuw, ...), ...
- This talk: test-time (adversarial) robustness of prediction methods (?2004–2013–...)

Adversarial robustness

- Let $f(x, \hat{\theta})$ be a predictor (classifier, regression function) estimated/learned from data
- For a new test datapoint x, we predict $\hat{y}(x) = f(x, \hat{\theta})$
- Suppose an adversary perturbs $x \to x' = x + \delta$ for some "small" adversarially chosen δ; forming adversarial example
- Want prediction $\hat{y}(x')$ to be robust/stable (not change much)

Adversarial robustness

- Let $f(x, \hat{\theta})$ be a predictor (classifier, regression function) estimated/learned from data
- For a new test datapoint x, we predict $\hat{y}(x) = f(x, \hat{\theta})$
- Suppose an adversary perturbs $x \to x' = x + \delta$ for some "small" adversarially chosen δ; forming adversarial example
- Want prediction $\hat{y}(x')$ to be robust/stable (not change much)

Modern methods are not robust

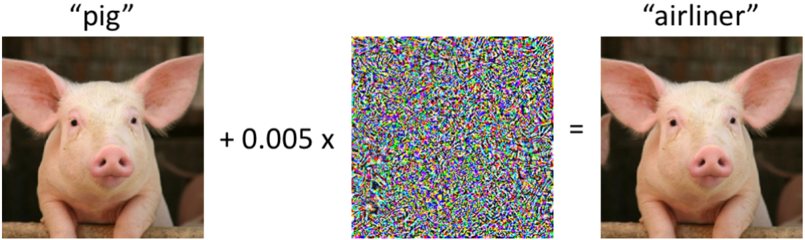


Figure: `gradientscience.org/intro_adversarial/`, by Madry & Schmidt. See works by Biggio et al (2013), Szegedy et al (2013), ...

- Key challenge in deploying learning algorithms for real problems

Modern methods are not robust: Adversarial patch

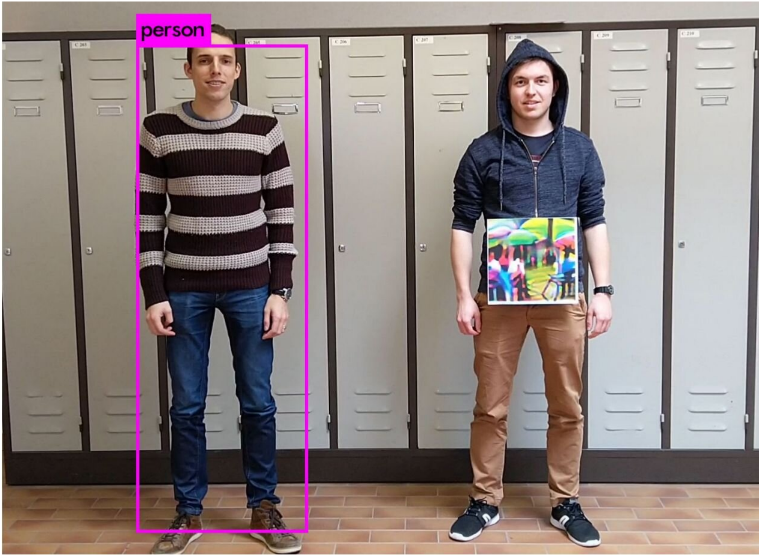
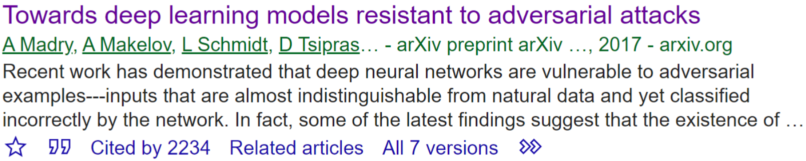


Figure: Thys et al. Fooling automated surveillance cameras: adversarial patches to attack person detection, arXiv:1904.08653

Unprecedented interest



Towards deep learning models resistant to adversarial attacks

A Madry, A Makelov, L Schmidt, D Tsipras… - arXiv preprint arXiv …, 2017 - arxiv.org

Recent work has demonstrated that deep neural networks are vulnerable to adversarial examples---inputs that are almost indistinguishable from natural data and yet classified incorrectly by the network. In fact, some of the latest findings suggest that the existence of …

Key questions

- What causes adversarial examples? e.g., model class (deep nets), training methods, data properties?
- How is this related to human learning? to child development?
- How to use this?

Overview

Overview

Theoretical explanations proposed

What causes adversarial examples?

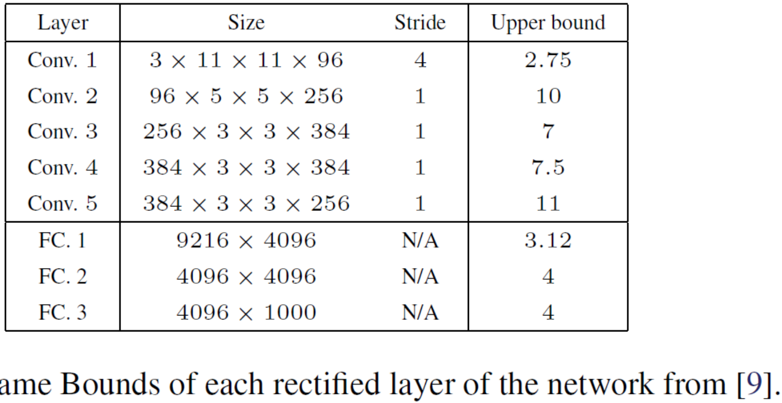
Theory	Data	Archit & Train	Experimental test	Notes
Lipschitz constant	Large cond num of weight mx		is adv ex aligned with top sing vec? appears with large sv?	
High dimensionality	truly high-dim, "uniform", deterministic	nonzero pop err	shouldn't work in low-dim?	real data low-dim?
Low-dim manifold	Unknown low-dim manifold		where are the adv ex in gen-ve model?	untestable in practice?
Oscillation		"dense" level sets; implied by interpo+Bayes>0	how close is random img to dog?	interpo+Bayes >0 not enough in high dim;
Non-Robust features	exist non-rob features $\mathbb{E} yf(x) > 0$, $\mathbb{E} \inf_\delta yf(x+\delta) = 0$	use non-rob features		somewhat circular

Table: Summary of theories to explain adversarial examples. For each approach, we list the assumptions on the data, the architecture and training. We also list possible experiments to test it, and other notes/limitations.

Lipschitz constant

- data $x \to$ prediction $f(x)$;
  $f(x+\delta)$ very different from $f(x)$
- model: Large Lipschitz constant of f — controlled by operator norms of weight matrices
- Szegedy et al. (2013): "Independently of generalisation properties, adv examples show that there exist small perturbations of the input that produce large perturbations of the output"

Lipschitz constant



Layer	Size	Stride	Upper bound
Conv. 1	$3 \times 11 \times 11 \times 96$	4	2.75
Conv. 2	$96 \times 5 \times 5 \times 256$	1	10
Conv. 3	$256 \times 3 \times 3 \times 384$	1	7
Conv. 4	$384 \times 3 \times 3 \times 384$	1	7.5
Conv. 5	$384 \times 3 \times 3 \times 256$	1	11
FC. 1	9216×4096	N/A	3.12
FC. 2	4096×4096	N/A	4
FC. 3	4096×1000	N/A	4

Table 5: Frame Bounds of each rectified layer of the network from [9].

Figure: Szegedy et al. (2013): Upper bounds on AlexNet Lipschitz constants

Lipschitz constant

- doesn’t take into account that we need to move examples across boundary (so only a sort of directional derivative; only for points near boundary matters)
- connection to human learning/development: erratic behavior/lack of focus/attending to artifacts?

High dimensionality

- data: truly high-dim, "uniform"
- nonzero test error
- concentration of measure: neighborhood of error set grows fast

High dimensionality

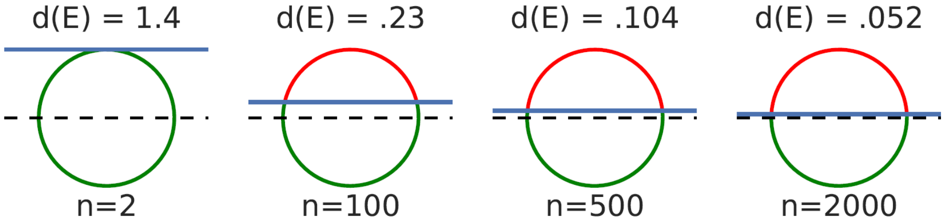
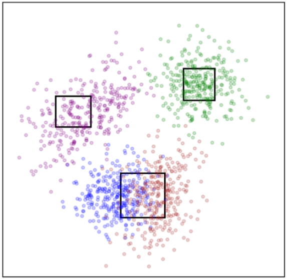
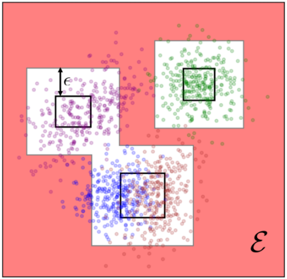


Figure: Unit sphere data. Region of 1% error. Gilmer et al. (2018).
$n = 2000$: 50% of points are within distance 0.05 from 1%

High dimensionality - empirically



cover the densest area using rectangles



expand the rectangles and treat the complement as the error region

Datasets	Risk Constraint (α)	Max Perturbation	Lower Bound on Adversarial Risk	Attack Success Rate for State-of-the-art Defenses
MNIST	0.01	$\ell_\infty \leq 0.3$	7.2%	10.7% [Madry+, 2018]
MNIST	0.01	$\ell_2 \leq 1.5$	2.1%	20.0% [Schott+, 2019]
CIFAR-10	0.05	$\ell_\infty \leq 8/255$	18.1%	52.9% [Madry+, 2019]

Figure: Empirically Measuring Concentration. Mahloujifar et al. (2019); worst-case dist

High dimensionality

- real data lies on low-dim manifold in a high dim space?
- connection to human learning: huge space of highly complex objects, many components?

Low-dim manifold

- data: on unknown low-dim manifold in high dim
- adv examples: off the manifold

Low-dim manifold

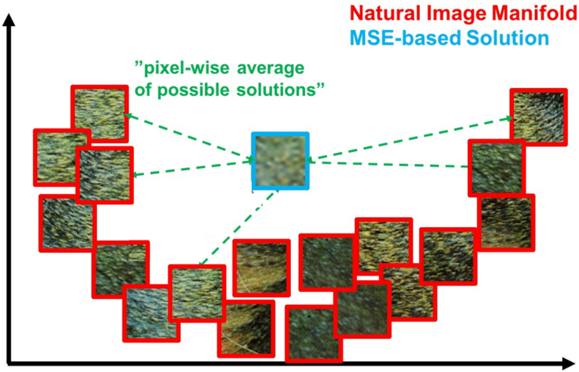
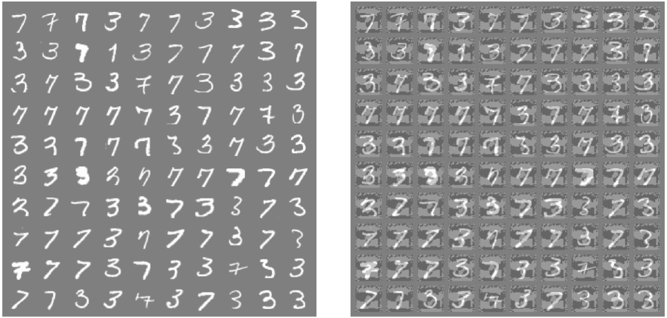


Figure: Ledig et al, 2016

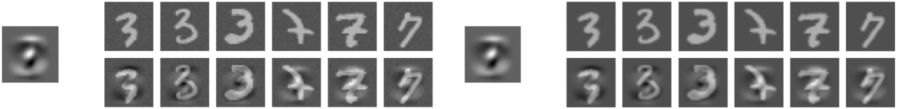
Low-dim manifold



(b) Left: original images from MNIST. Right: adversarial examples with logistic regression.

Figure: Tanay & Griffin, 2016

Low-dim manifold



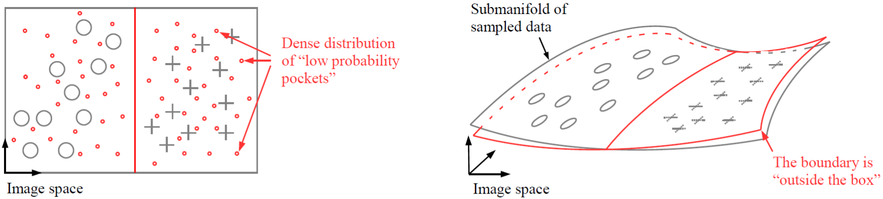
(a) 3s vs 7s MNIST problem with an image size of 28×28. Left: weight vector defined by linear SVM. Right: example digits (top) and their adversarial examples (bottom).

(b) 3s vs 7s MNIST problem with an image size of 200×200. Left: weight vector defined by linear SVM. Right: the same example digits (top) and their adversarial examples (bottom).

Figure 2: Increasing the dimensionality of the problem does not make the phenomenon of adversarial examples worse. Whether the image size is 28×28 or 200×200, the weight vector found by linear SVM looks very similar to the one found by logistic regression in (Goodfellow et al., 2014). The two SVM models have an error rate of 2.7%[2]. The magnitude ϵ of the perturbations has been chosen in both cases such that 99% of the digits in the test set are misclassified ($\epsilon_{28} = 4.6, \epsilon_{200} = 31. \approx \epsilon_{28} \times 200/28$)

Figure: Tanay & Griffin, 2016

Low-dim manifold



(a) The solution proposed in (Szegedy et al., 2013). Adversarial examples are possible because the image space is densely filled with low probability adversarial pockets.

(b) The solution we propose. Adversarial examples are possible because the class boundary extends beyond the submanifold of sample data and can be — under certain circumstances — lying close to it.

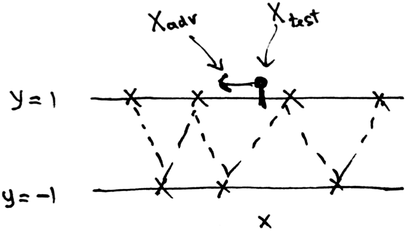
Figure: Tanay & Griffin, 2016

Low-dim manifold

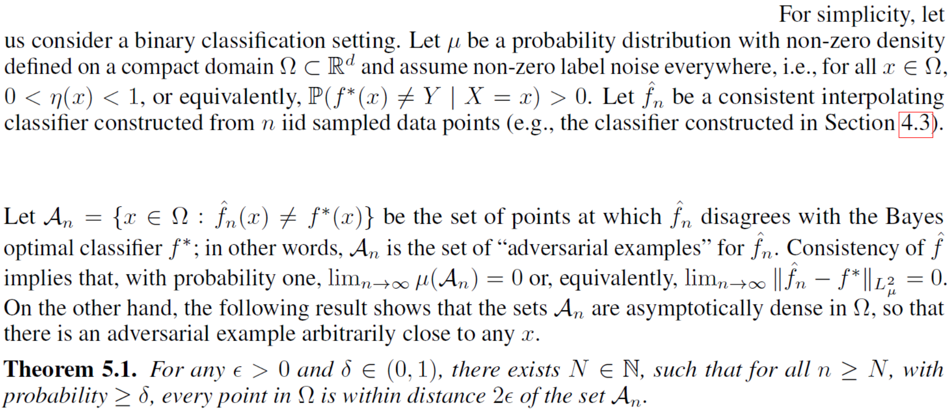
- testing this is hard: don't know manifold
- where are the adv ex in generative model?
- how to use this? related idea: unlabeled data helps with adversarial robustness (can learn manifold); robust self-training (Carmon et al., 2019)
- missing: properties of classifier; principled use
- connection to human learning: ??

Oscillation

- classifier has "dense" classification regions
- Szegedy et al. (2013): "set of adv. negatives is dense (much like the rational numbers), and so it is found virtually near every test case"
- implied by interpolation + Bayes risk>0 (prop. of deep net training + prop. of data)



Oscillation



For simplicity, let us consider a binary classification setting. Let μ be a probability distribution with non-zero density defined on a compact domain $\Omega \subset \mathbb{R}^d$ and assume non-zero label noise everywhere, i.e., for all $x \in \Omega$, $0 < \eta(x) < 1$, or equivalently, $\mathbb{P}(f^*(x) \neq Y \mid X = x) > 0$. Let $\hat{f}_n$ be a consistent interpolating classifier constructed from n iid sampled data points (e.g., the classifier constructed in Section 4.3).

Let $\mathcal{A}_n = \{x \in \Omega : \hat{f}_n(x) \neq f^*(x)\}$ be the set of points at which $\hat{f}_n$ disagrees with the Bayes optimal classifier f^*; in other words, $\mathcal{A}_n$ is the set of "adversarial examples" for $\hat{f}_n$. Consistency of $\hat{f}$ implies that, with probability one, $\lim_{n\to\infty} \mu(\mathcal{A}_n) = 0$ or, equivalently, $\lim_{n\to\infty} \|\hat{f}_n - f^*\|_{L^2_\mu} = 0$. On the other hand, the following result shows that the sets $\mathcal{A}_n$ are asymptotically dense in Ω, so that there is an adversarial example arbitrarily close to any x.

Theorem 5.1. *For any $\epsilon > 0$ and $\delta \in (0, 1)$, there exists $N \in \mathbb{N}$, such that for all $n \geq N$, with probability $\geq \delta$, every point in Ω is within distance 2ϵ of the set $\mathcal{A}_n$.*

Figure: Belkin et al. (2018)

Oscillation

- test: how close is random img to dog?
- use: regularize oscillations
- one approach is TRADES (Zhang et al., 2019); add penalty $f(x)f(x')$
- others: stability training (Gaussian), adversarial logit pairing
- connection to human learning: low-resolution, high-noise settings, each object is ambiguous?

Non-Robust features

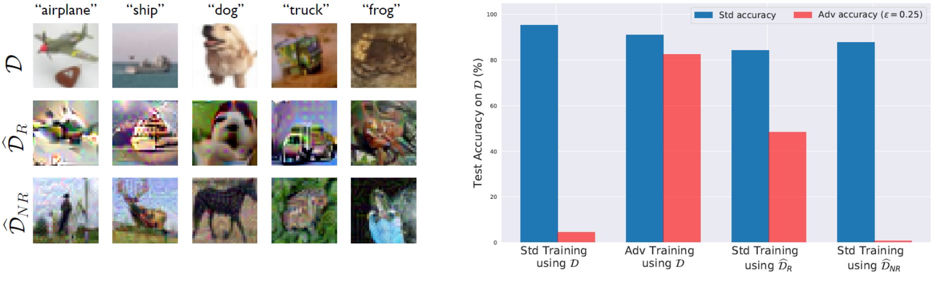
- Data: exist non-rob features g

$$\mathbb{E} y g(x) > 0,$$

$$\mathbb{E} \inf_{\|\delta\| \leq \varepsilon} y g(x + \delta) = 0$$

- model: use non-rob features
- somewhat circular (i.e., true by definition)

Non-Robust features



(a) (b)

Figure 2: **Left**: Random samples from our variants of the CIFAR-10 [Kri09] training set: the original training set; the *robust training set* $\widehat{\mathcal{D}}_R$, restricted to features used by a robust model; and the *non-robust training set* $\widehat{\mathcal{D}}_{NR}$, restricted to features relevant to a standard model (labels appear incorrect to humans). **Right**: Standard and robust accuracy on the CIFAR-10 test set ($\mathcal{D}$) for models trained with: (i) standard training (on $\mathcal{D}$) ; (ii) standard training on $\widehat{\mathcal{D}}_{NR}$; (iii) adversarial training (on $\mathcal{D}$); and (iv) standard training on $\widehat{\mathcal{D}}_R$. Models trained on $\widehat{\mathcal{D}}_R$ and $\widehat{\mathcal{D}}_{NR}$ reflect the original models used to create them: notably, standard training on $\widehat{\mathcal{D}}_R$ yields nontrivial robust accuracy. Results for Restricted-ImageNet [Tsi+19] are in D.8 Figure 12.

Figure: Ilyas et al. (2019): $\hat{D}_R$ projects data into rob features

Non-Robust features

- how to use?
- connection to human learning: complex settings with nuisances?

Others

- computational constraints
- sample needs to be much larger than for standard training
- class imbalance
- need overparametrization
- see also Robustness Session at `c3dti.ai/events/workshops/foundations-of-deep-learning/`, YouTube video `www.youtube.com/watch?v=p4JpfmW5PNQ`

Summary

- Adversarial robustness: many proposed explanations
- How is this related to human learning? to child development?
- How to use this?
- Thanks!

References I

M. Belkin, D. J. Hsu, and P. Mitra. Overfitting or perfect fitting? risk bounds for classification and regression rules that interpolate. In *Advances in Neural Information Processing Systems 31*, pages 2300–2311. Curran Associates, Inc., 2018.

Y. Carmon, A. Raghunathan, L. Schmidt, J. C. Duchi, and P. S. Liang. Unlabeled data improves adversarial robustness. In *Advances in Neural Information Processing Systems*, pages 11192–11203, 2019.

J. Gilmer, L. Metz, F. Faghri, S. S. Schoenholz, M. Raghu, M. Wattenberg, and I. Goodfellow. The relationship between high-dimensional geometry and adversarial examples, 2018. URL `http://arxiv.org/abs/1801.02774v3`.

A. Ilyas, S. Santurkar, D. Tsipras, L. Engstrom, B. Tran, and A. Madry. Adversarial examples are not bugs, they are features. In *Advances in Neural Information Processing Systems*, pages 125–136, 2019.

S. Mahloujifar, X. Zhang, M. Mahmoody, and D. Evans. Empirically measuring concentration: Fundamental limits on intrinsic robustness. In *Advances in Neural Information Processing Systems*, pages 5210–5221, 2019.

References II

C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

H. Zhang, Y. Yu, J. Jiao, E. P. Xing, L. E. Ghaoui, and M. I. Jordan. Theoretically principled trade-off between robustness and accuracy. *arXiv preprint arXiv:1901.08573*, 2019.