Robustness and geometry of deep neural networks

Alhussein Fawzi
DeepMind

May 23rd 2019
The Mathematics of Deep Learning and Data Science
University of Cambridge

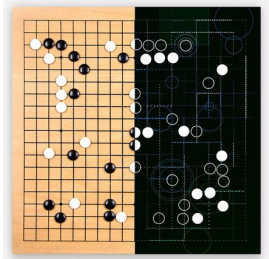
Recent advances in machine learning

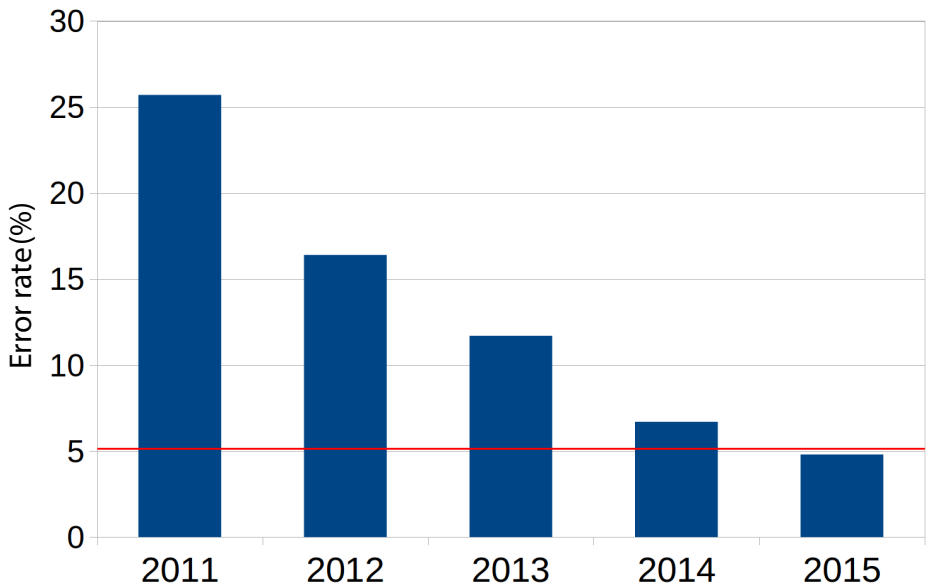
GT: horse cart
1: horse cart
2: minibus
3: oxcart
4: stretcher
5: half track

GT: birdhouse
1: birdhouse
2: sliding door
3: window screen
4: mailbox
5: pot

"man in black shirt is playing guitar."

"construction worker in orange safety vest is working on road."





He et., al., "Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification", 2015
Karpathy et., al, "Automated Image Captioning with ConvNets and Recurrent Nets"
LeCun et. al.,, "Deep Learning", 2015
DeepMind https://deepmind.com/research/alphago/

Robustness of classifiers to perturbations

- In real world environments, images undergo perturbations.

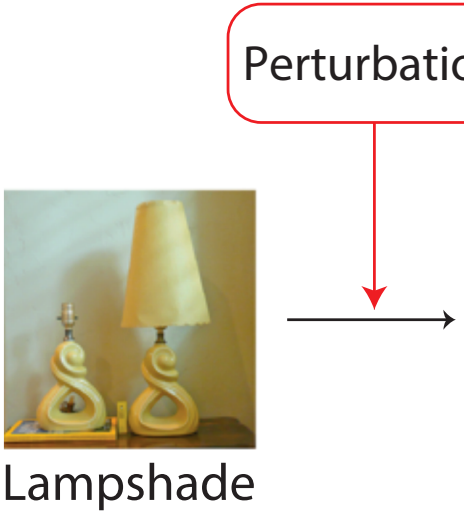
Robustness of classifiers to perturbations

- In real world environments, images undergo perturbations.

Lampshade

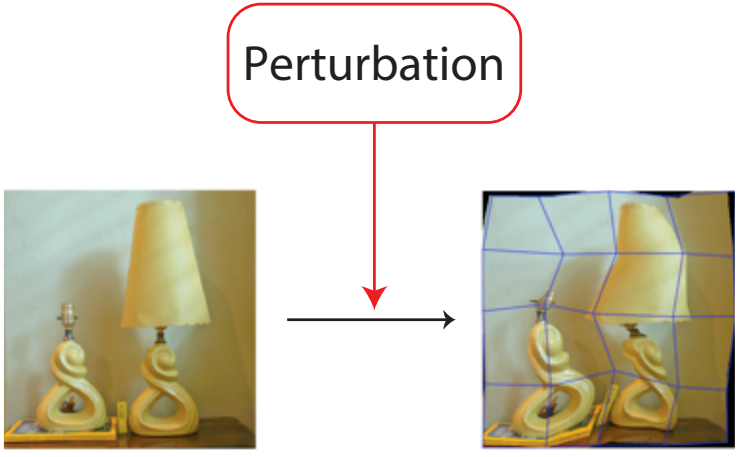
Robustness of classifiers to perturbations

- In real world environments, images undergo perturbations.



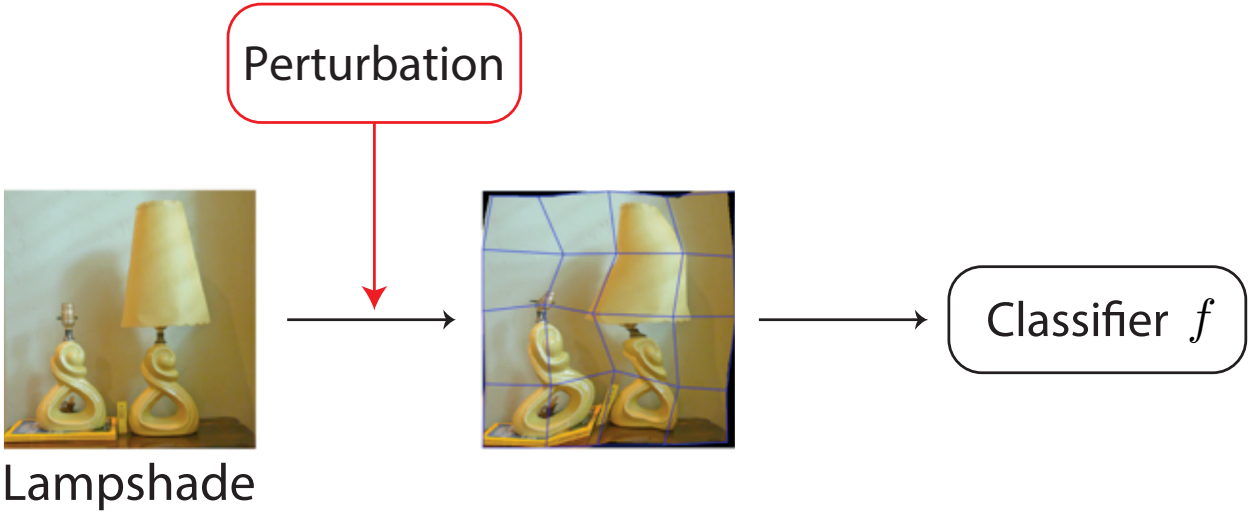
Robustness of classifiers to perturbations

- In real world environments, images undergo perturbations.



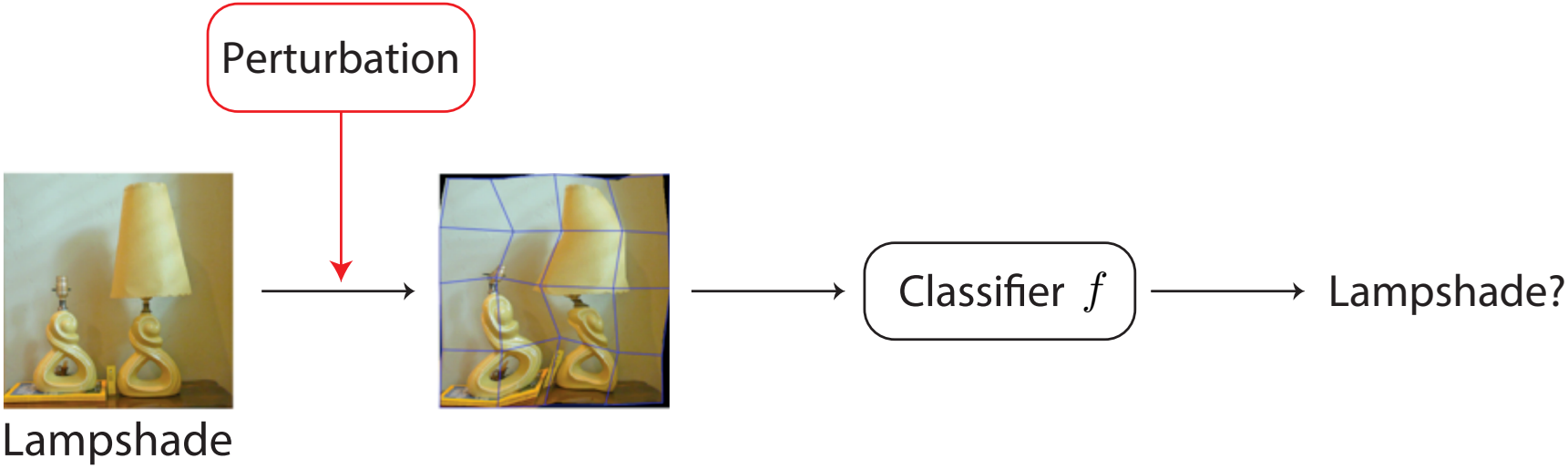
Robustness of classifiers to perturbations

- In real world environments, images undergo perturbations.



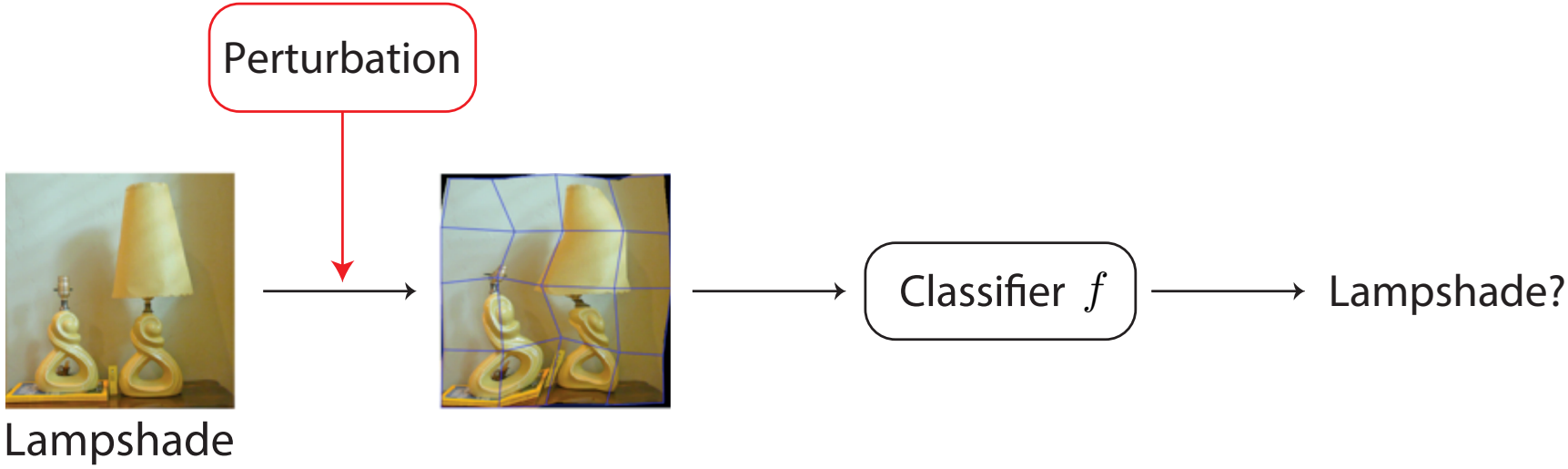
Robustness of classifiers to perturbations

- In real world environments, images undergo perturbations.



Robustness of classifiers to perturbations

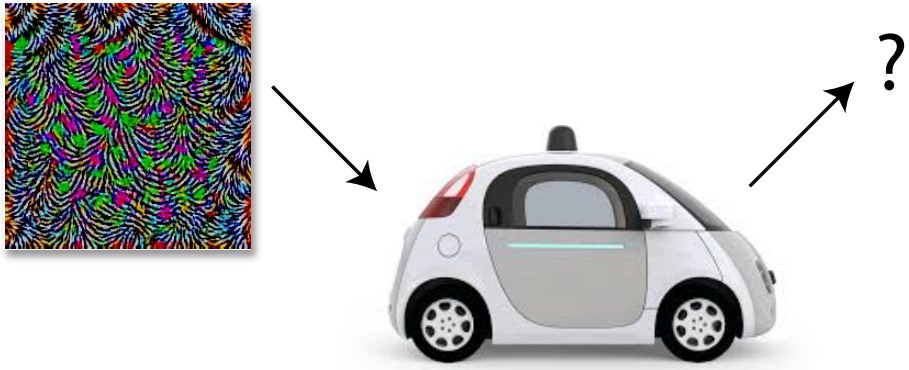
- In real world environments, images undergo perturbations.



- Broad range of perturbations
  - Adversarial perturbations [Szegedy et. al. ICLR 2014], [Biggio et. al., PKDD 2013], ...
  - Random noise [Fawzi et. al., NIPS 2016], [Franceschi et. al., AISTATS 2018]
  - Structured nuisances (geometric transformations [Bruna et. al., TPAMI 2013], [Jaderberg et. al., NIPS 2015] , occlusions [Sharif et. al., CCS 2016] , etc...).

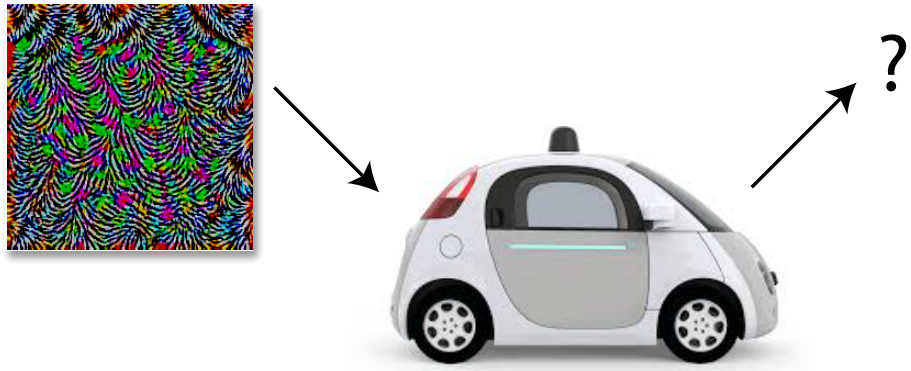
Robustness of classifiers to perturbations (Cont'd)

- Safety of machine learning systems

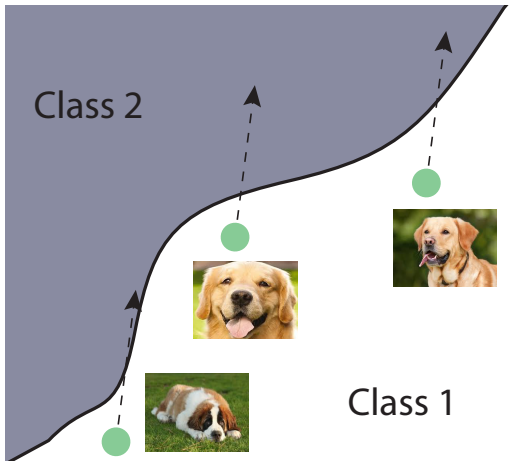


Robustness of classifiers to perturbations (Cont'd)

- Safety of machine learning systems



- Better understanding of the geometry of state-of-the-art classifiers.



Talk outline

1. Fooling classifiers is easy: vulnerability to different perturbations.
2. Improving the robustness (i.e., "defending") is difficult.
3. Geometric analysis of a successful defense: adversarial training.

Adversarial perturbations

- State-of-the-art deep neural networks have been shown to be surprisingly unstable to *adversarial* perturbations.

Adversarial perturbations

- State-of-the-art deep neural networks have been shown to be surprisingly unstable to *adversarial* perturbations.

School bus

Adversarial perturbations

- State-of-the-art deep neural networks have been shown to be surprisingly unstable to *adversarial* perturbations.

School bus

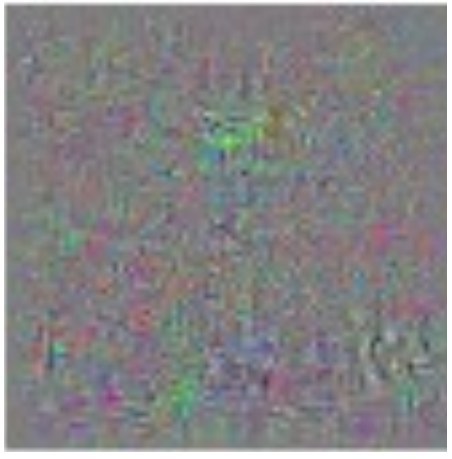
Ostrich

Adversarial perturbations

- State-of-the-art deep neural networks have been shown to be surprisingly unstable to *adversarial* perturbations.

School bus

Perturbation



Ostrich

Figure from [Szegedy et. al., ICLR 2014].

Adversarial perturbations

- State-of-the-art deep neural networks have been shown to be surprisingly unstable to *adversarial* perturbations.

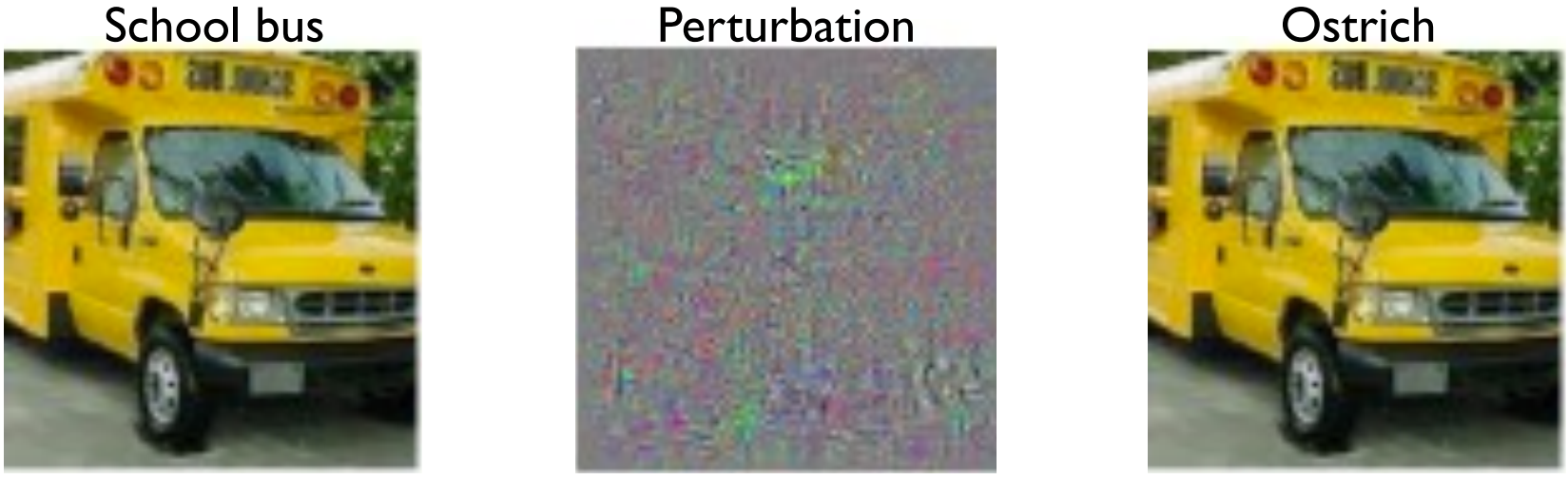
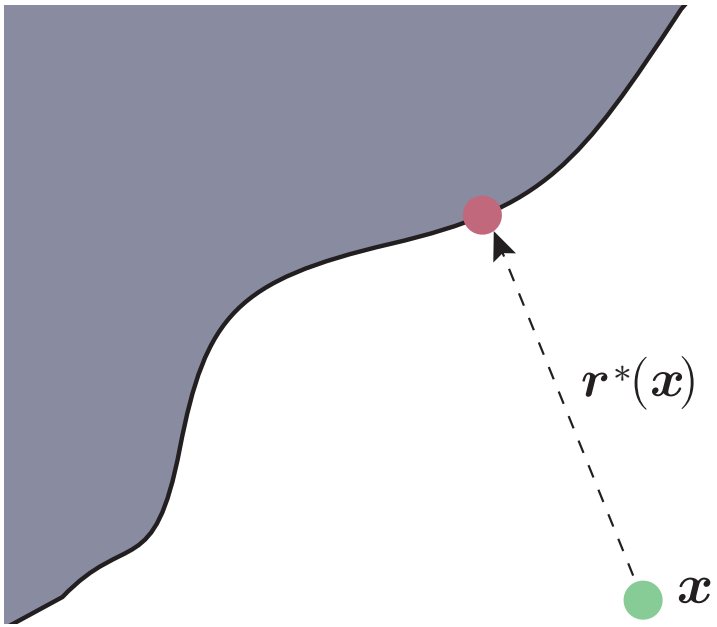


Figure from [Szegedy et. al., ICLR 2014].

- Adversarial examples are found by seeking the minimal perturbation (in the ℓ_2 sense) that switches the label of the classifier.

Adversarial perturbations

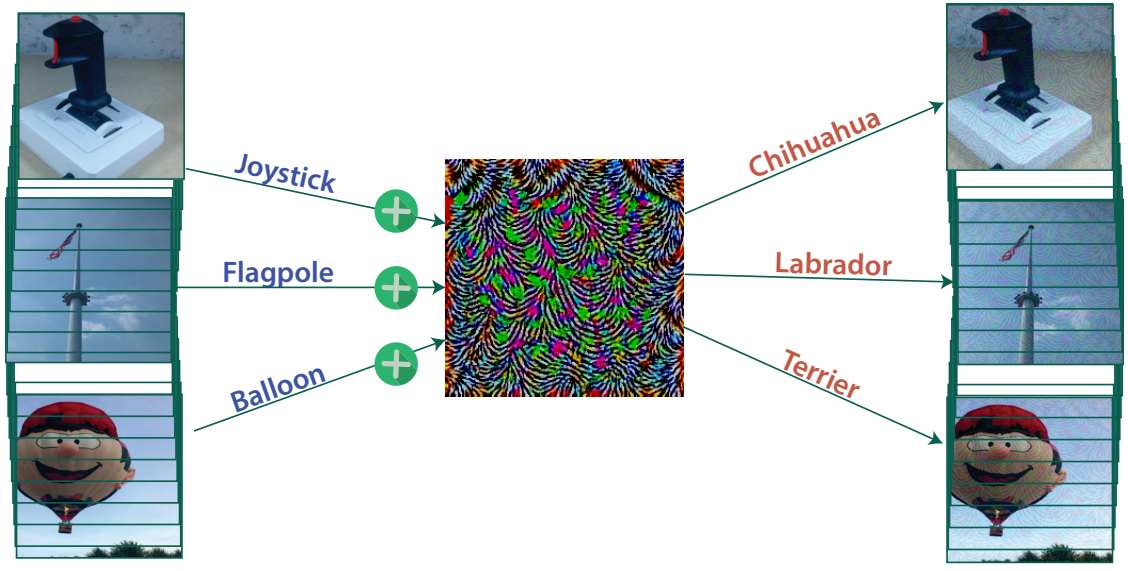
Robustness to adversarial noise



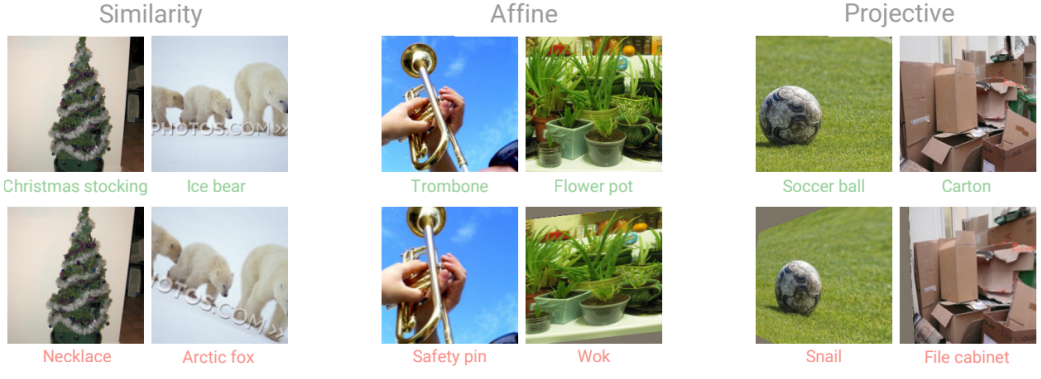
$$r^*(x) = \min_r \|r\|_2 \text{ subject to } f(x+r) \neq f(x).$$

Other types of adversarial perturbations

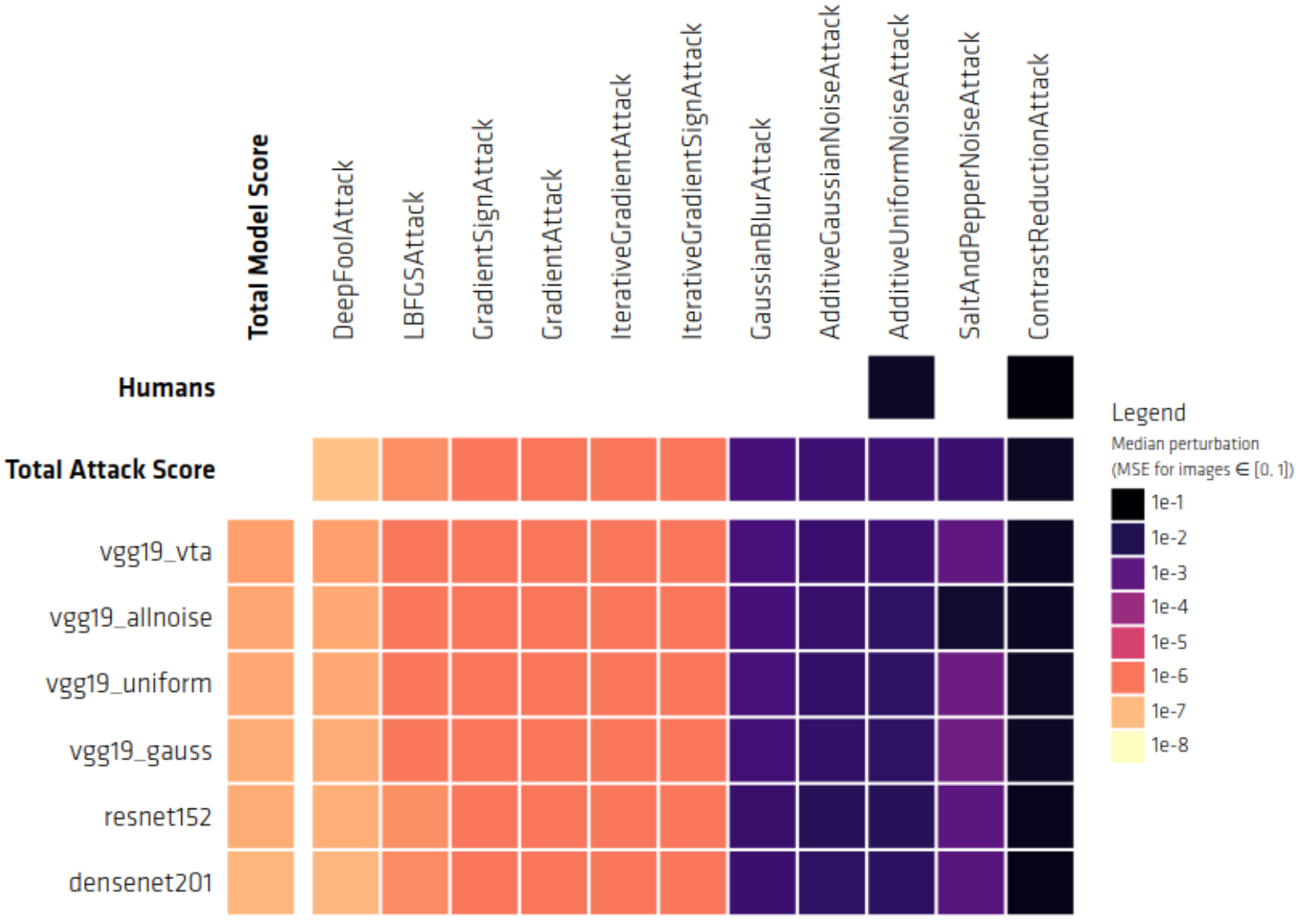
- Universal perturbations [Moosavi-Dezfooli et. al., 2017]



- Geometric transformations [Fawzi et. al., 2015, Moosavi-Dezfooli et. al., 2018, Xiao et al., 2018]



Finding adversarial perturbations is easy...



http://robust.vision

... but designing defense mechanisms is hard!

Despite the huge number of proposed defenses, state-of-the-art classifiers are still vulnerable to small perturbations.

Adversarial Risk and the Dangers of Evaluating Against Weak Attacks

Jonathan Uesato [1] Brendan O'Donoghue [1] Aaron van den Oord [1] Pushmeet Kohli [1]

Abstract

This paper investigates recently proposed approaches for defending against adversarial examples and evaluating adversarial robustness. We motivate *adversarial risk* as an objective for achieving models robust to worst-case inputs. We then frame commonly used attacks and evaluation metrics as defining a tractable surrogate objective to the true adversarial risk. This suggests that models may optimize this surrogate rather than the true adversarial risk. We formalize this notion as *obscurity to an adversary*, and develop tools and heuristics for identifying obscured models and designing transparent models. We demonstrate that this is a significant problem in practice by repurposing gradient-free optimization techniques into adversarial attacks, which we use to decrease the accuracy of several recently proposed defenses to near zero. Our hope is that our formulations and results will help researchers to develop more powerful defenses.

1. Introduction

Deep learning has revolutionized the field of machine learning and led to substantial improvements in many challenging problems such as image understanding (He et al., 2016), speech recognition (Graves et al., 2013), and automatic game playing (Mnih et al., 2015). Despite these remarkable successes, we have seen some intriguing and troubling properties in the behaviour of these models.

Researchers have demonstrated that certain small perturbations to the input can make neural networks produce extremely bad results (Szegedy et al., 2013; Jia and Liang, 2017) For instance, in the case of image classification, imperceptible perturbations can lead to the resulting images being misclassified to completely different object categories with high confidence (Szegedy et al., 2013). These so-called *adversarial* examples can be computed relatively easily by using optimization methods (referred to as adversarial attacks) to find perturbations that maximize the loss of the network (Goodfellow et al., 2014; Carlini and Wagner, 2017b). Later work showed that this phenomena is not unique to image classification, and appears across different model architectures, as well as in other machine learning algorithms (Papernot et al., 2016; 2017).

The emergence of adversarial examples and the increasing deployment of machine learning models in real world production systems has motivated extensive research on developing models that can defend against adversarial attacks (Warde-Farley and Goodfellow, 2016; Yuan et al., 2017). Using a variety of approaches, these works have shown that neural models can be developed that are robust against commonly used attacks (Guo et al., 2017; Song et al., 2017; Xie et al., 2017; Liao et al., 2017). However, a key question remains unanswered: *Are these models free from any adversarial examples or are they simply robust to current attack methods?*

In this paper, we formalize the intuition that in settings with the potential for catastrophic failures, minimizing expected risk may produce models with very poor worst-case performance. This motivates the study of the adversarial risk as a measure of the model's performance on worst-case inputs. However, the exact adversarial risk is computationally intractable to evaluate, let alone optimize. We thus frame commonly used attacks and adversarial evaluation metrics as defining a tractable surrogate objective to the true adversarial risk. We hypothesize that many defenses achieve robustness through *obscurity*, *i.e.*, the defenses work by exploiting weaknesses of certain attacks and do not eliminate all adversarial inputs.

One of the key contributions of this paper is to experimentally validate the 'security by obscurity' nature of recently proposed defense methods. Specifically, we show that by using a more powerful attack method (one better able to maximize the true adversarial risk), we can dramatically reduce the performance of these defenses.

To summarize, the key contributions of this paper are:

- Formulation of adversarial attacks and defenses as optimizing *surrogates* of the true adversarial risk

[1]DeepMind. Correspondence to: Jonathan Uesato <juesato@google.com>.

Proceedings of the 35th International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018.

Obfuscated Gradients Give a False Sense of Security: Circumventing Defenses to Adversarial Examples

Anish Athalye [*1] Nicholas Carlini [*2] David Wagner [2]

Abstract

We identify obfuscated gradients, a kind of gradient masking, as a phenomenon that leads to a false sense of security in defenses against adversarial examples. While defenses that cause obfuscated gradients appear to defeat iterative optimization-based attacks, we find defenses relying on this effect can be circumvented. We describe characteristic behaviors of defenses exhibiting the effect, and for each of the three types of obfuscated gradients we discover, we develop attack techniques to overcome it. In a case study, examining non-certified white-box-secure defenses at ICLR 2018, we find obfuscated gradients are a common occurrence, with 7 of 9 defenses relying on obfuscated gradients. Our new attacks successfully circumvent 6 completely, and 1 partially, in the original threat model each paper considers.

1. Introduction

In response to the susceptibility of neural networks to adversarial examples (Szegedy et al., 2013; Biggio et al., 2013), there has been significant interest recently in constructing defenses to increase the robustness of neural networks. While progress has been made in understanding and defending against adversarial examples in the white-box setting, where the adversary has full access to the network, a complete solution has not yet been found.

As benchmarking against iterative optimization-based attacks (e.g., Kurakin et al. (2016a); Madry et al. (2018); Carlini & Wagner (2017c)) has become standard practice in evaluating defenses, new defenses have arisen that appear to be robust against these powerful optimization-based attacks.

We identify one common reason why many defenses provide apparent robustness against iterative optimization attacks: *obfuscated gradients*, a term we define as a special case of gradient masking (Papernot et al., 2017). Without a good gradient, where following the gradient does not successfully optimize the loss, iterative optimization-based methods cannot succeed. We identify three types of obfuscated gradients: *shattered gradients* are nonexistent or incorrect gradients caused either intentionally through non-differentiable operations or unintentionally through numerical instability; *stochastic gradients* depend on test-time randomness; and *vanishing/exploding gradients* in very deep computation result in an unusable gradient.

We propose new techniques to overcome obfuscated gradients caused by these three phenomena. We address gradient shattering with a new attack technique we call Backward Pass Differentiable Approximation, where we approximate derivatives by computing the forward pass normally and computing the backward pass using a differentiable approximation of the function. We compute gradients of randomized defenses by applying Expectation Over Transformation (Athalye et al., 2017). We solve vanishing/exploding gradients through reparameterization and optimize over a space where gradients do not explode/vanish.

To investigate the prevalence of obfuscated gradients and understand the applicability of these attack techniques, we use as a case study the ICLR 2018 non-certified defenses that claim white-box robustness. We find that obfuscated gradients are a common occurrence, with 7 of 9 defenses relying on this phenomenon. Applying the new attack techniques we develop, we overcome obfuscated gradients and circumvent 6 of them completely, and 1 partially, under the original threat model of each paper. Along with this, we offer an analysis of the evaluations performed in the papers.

Additionally, we hope to provide researchers with a common baseline of knowledge, description of attack techniques, and common evaluation pitfalls, so that future defenses can avoid falling vulnerable to these same attack approaches.

To promote reproducible research, we release our re-implementation of each of these defenses, along with implementations of our attacks for each. [1]

[*]Equal contribution [1]Massachusetts Institute of Technology [2]University of California, Berkeley. Correspondence to: Anish Athalye <aathalye@mit.edu>, Nicholas Carlini <npc@berkeley.edu>.

Proceedings of the 35th International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018.

[1] https://github.com/anishathalye/obfuscated-gradients

... but designing defense mechanisms is hard!

Despite the huge number of proposed defenses, state-of-the-art classifiers are still vulnerable to small perturbations.

Adversarial Risk and the Dange

Jonathan Uesato [1] Brendan O'Donog

Abstract

This paper investigates recently proposed approaches for defending against adversarial examples and evaluating adversarial robustness. We motivate *adversarial risk* as an objective for achieving models robust to worst-case inputs. We then frame commonly used attacks and evaluation metrics as defining a tractable surrogate objective to the true adversarial risk. This suggests that models may optimize this surrogate rather than the true adversarial risk. We formalize this notion as *obscurity to an adversary*, and develop tools and heuristics for identifying obscured models and designing transparent models. We demonstrate that this is a significant problem in practice by repurposing gradient-free optimization techniques into adversarial attacks, which we use to decrease the accuracy of several recently proposed defenses to near zero. Our hope is that our formulations and results will help researchers to develop more powerful defenses.

1. Introduction

Deep learning has revolutionized the field of machine le ing and led to substantial improvements in many challe ing problems such as image understanding (He et al., 20 speech recognition (Graves et al., 2013), and autom game playing (Mnih et al., 2015). Despite these rem able successes, we have seen some intriguing and troubl properties in the behaviour of these models.

Researchers have demonstrated that certain small pertur tions to the input can make neural networks produce tremely bad results (Szegedy et al., 2013; Jia and Lia 2017) For instance, in the case of image classification, perceptible perturbations can lead to the resulting ima being misclassified to completely different object catego with high confidence (Szegedy et al., 2013). These so-cal

[1]DeepMind. Correspondence to: Jonathan Uesato <sato@google.com>.

Proceedings of the 35th International Conference on Mach... Learning, Stockholm, Sweden, PMLR 80, 2018. Copyright 2018 by the author(s).

arXiv:1802.05666v2 [cs.LG] 12 Jun 2018

arXiv:1711.08478v1 [cs.LG] 22 Nov 2017

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

Nicholas Carlini David Wagner
University of California, Berkeley

Abstract

MagNet and "Efficient Defenses..." were recently proposed as a defense to adversarial examples. We find that we can construct adversarial examples that defeat these defenses with only a slight increase in distortion.

1 Introduction

It is an open question how to train neural networks so they will be robust to adversarial examples [11]. Recently, three defenses have been proposed to make neural networks robust to adversarial examples:

- MagNet [8] was proposed as an approach to make neural networks robust against adversarial examples through two complementary approaches: adversarial examples near the data manifold are *reformed* to lie on the data manifold that are classified correctly, whereas adversarial examples far away from the data manifold are *detected* and rejected before classification. MagNet does not argue robustness in the white-box setting; rather, the authors argue that MagNet is robust in the grey-box setting where the adversary is aware the defense is in place, knows the parameters of the base classifier, but *not* the parameters of the defense.
- An efficient defense [12] was proposed to make neural networks more robust against adversarial examples by performing Gaussian data augmentation during training, and using the BReLU activation function. The authors do not claim perfect security, but claim this makes attacks visually detectable.
- Adversarial Perturbation Elimination GAN (APE-GAN) [10] is similar to MagNet, only adversarial examples are projected onto the data manifold using a Generative Adversarial Network (GAN) [2]. We did not set out to bypass this defense, but found it to be very similar to MagNet and so we analyze it too.

In this short paper, we demonstrate these three defenses are not effective on the MNIST [5] and CIFAR-10 [4] datasets. We show that we are able to bypass MagNet with greater than 99% success, and the latter two with 100%, with only a slight increase in distortion.

We defeat MagNet by making use of the *transferability* [11] property of adversarial examples: the adversary trains their own copy of the defense, constructs adversarial examples on their model, and supplies these adversarial examples to the defender. It turns out that these examples will also fool the defender's model.

We defeat "Efficient Defenses Against Adversarial Attack" and APE-GAN by showing that existing attack can defeat them with 100% success without modification. Adversarial examples are not more visually detectable than an undefended network.

2 Background

We assume familiarity with neural networks [5], adversarial examples [11], transferability [6], generating strong attacks against adversarial examples [1] and MagNet [8]. We briefly review the key details and notation.

Notation Let $F(x) = y$ be a neural network used for classification outputting a probability distribution. Call the second-to-last layer (the layer before the the softmax layer) Z, so that $F(x) = \text{softmax}(Z(x))$. Each output y_i corresponds to the predicted probability that the object x is labelled as class i. Let $C(x) = \arg\max_i \ F(x_i)$ correspond to the classification of x on F. In this paper we are concerned with neural networks used to classify images (on MNIST and CIFAR-10).

Adversarial examples [11] are instances x' that are very close to a normal instance x with respect to some distance metric (L_2 distance, in this paper), but where $C(x') = t$ for any target t chosen by the adversary.

We generate adversarial examples with Carlini and Wagner's L_2 attack algorithm [1]. Specifically, we solve

$$\text{minimize} \ \ \|x' - x\|_2^2 + c \cdot \ell(x')$$

1

- Formulation of adversarial attacks and defenses as optimizing *surrogates* of the true adversarial risk

by the author(s).

...uscated Gradients Give a False Sense of Security: ...rcumventing Defenses to Adversarial Examples

Anish Athalye [*1] Nicholas Carlini [*2] David Wagner [2]

...stract

...l gradients, a kind of gradi... ...omenon that leads to a false ...efenses against adversarial ...nses that cause obfuscated ...feat iterative optimization... ...d defenses relying on this ...ented. We describe charac... ...fenses exhibiting the effect, ...ee types of obfuscated gra... ...develop attack techniques ...ase study, examining non... ...ure defenses at ICLR 2018, ...dients are a common occur... ...nses relying on obfuscated ...tacks successfully circum... ...l 1 partially, in the original ...er considers.

...ility of neural networks to adver... ...t al., 2013; Biggio et al., 2013), ...terest recently in constructing de... ...tness of neural networks. While ...n understanding and defending ...s in the white-box setting, where ...ess to the network, a complete ...ound.

...iterative optimization-based at... ...(2016a); Madry et al. (2018); ...has become standard practice in ...fenses have arisen that appear to ...erful optimization-based attacks.

...ason why many defenses provide

...ssachusetts Institute of Technol... ...nia, Berkeley. Correspondence ...lye@mit.edu>, Nicholas Carlini

...rnational Conference on Machine ..., PMLR 80, 2018. Copyright 2018

apparent robustness against iterative optimization attacks: *obfuscated gradients*, a term we define as a special case of gradient masking (Papernot et al., 2017). Without a good gradient, where following the gradient does not successfully optimize the loss, iterative optimization-based methods cannot succeed. We identify three types of obfuscated gradients: *shattered gradients* are nonexistent or incorrect gradients caused either intentionally through non-differentiable operations or unintentionally through numerical instability; *stochastic gradients* depend on test-time randomness; and *vanishing/exploding gradients* in very deep computation result in an unusable gradient.

We propose new techniques to overcome obfuscated gradients caused by these three phenomena. We address gradient shattering with a new attack technique we call Backward Pass Differentiable Approximation, where we approximate derivatives by computing the forward pass normally and computing the backward pass using a differentiable approximation of the function. We compute gradients of randomized defenses by applying Expectation Over Transformation (Athalye et al., 2017). We solve vanishing/exploding gradients through reparameterization and optimize over a space where gradients do not explode/vanish.

To investigate the prevalence of obfuscated gradients and understand the applicability of these attack techniques, we use as a case study the ICLR 2018 non-certified defenses that claim white-box robustness. We find that obfuscated gradients are a common occurrence, with 7 of 9 defenses relying on this phenomenon. Applying the new attack techniques we develop, we overcome obfuscated gradients and circumvent 6 of them completely, and 1 partially, under the original threat model of each paper. Along with this, we offer an analysis of the evaluations performed in the papers.

Additionally, we hope to provide researchers with a common baseline of knowledge, description of attack techniques, and common evaluation pitfalls, so that future defenses can avoid falling vulnerable to these same attack approaches.

To promote reproducible research, we release our re-implementation of each of these defenses, along with implementations of our attacks for each. [1]

[1] https://github.com/anishathalye/obfuscated-gradients

... but designing defense mechanisms is hard!

Despite the huge number of proposed defenses, state-of-the-art classifiers are still vulnerable to small perturbations.

Adversarial Risk and the Dange

Jonathan Uesato [1] Brendan O'Donog

Abstract

This paper investigates recently proposed approaches for defending against adversarial examples and evaluating adversarial robustness. We motivate *adversarial risk* as an objective for achieving models robust to worst-case inputs. We then frame commonly used attacks and evaluation metrics as defining a tractable surrogate objective to the true adversarial risk. This suggests that models may optimize this surrogate rather than the true adversarial risk. We formalize this notion as *obscurity to an adversary*, and develop tools and heuristics for identifying obscured models and designing transparent models. We demonstrate that this is a significant problem in practice by repurposing gradient-free optimization techniques into adversarial attacks, which we use to decrease the accuracy of several recently proposed defenses to near zero. Our hope is that our formulations and results will help researchers to develop more powerful defenses.

1. Introduction

Deep learning has revolutionized the field of machine le ing and led to substantial improvements in many challe ing problems such as image understanding (He et al., 20 speech recognition (Graves et al., 2013), and automa game playing (Mnih et al., 2015). Despite these rema able successes, we have seen some intriguing and troubl properties in the behaviour of these models.

Researchers have demonstrated that certain small pertu tions to the input can make neural networks produce tremely bad results (Szegedy et al., 2013; Jia and Lia 2017) For instance, in the case of image classification, perceptible perturbations can lead to the resulting ima being misclassified to completely different object catego with high confidence (Szegedy et al., 2013). These so-cal

[1]DeepMind. Correspondence to: Jonathan Uesato <sato@google.com>.

Proceedings of the 35th International Conference on Mach..... Learning, Stockholm, Sweden, PMLR 80, 2018.

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

Nicholas Carlini David Wagner
University of California, Berkeley

Abstract

MagNet and "Efficient Defenses..." were recently proposed as a defense to adversarial examples. We find that we can construct adversarial examples that defeat these defenses with only a slight increase in distortion.

In this short paper, we demonstrate these fenses are not effective on the MNIST [5] and [4] datasets. We show that we are able to by Net with greater than 99% success, and the with 100%, with only a slight increase in disto

We defeat MagNet by making use of the *tra ity* [11] property of adversarial examples: the trains their own copy of the defense, constructs ial examples on their model, and supplies th sarial examples to the defender. It turns out examples will also fool the defender's model.

We defeat "Efficient Defenses Against Adve tack" and APE-GAN by showing that existing defeat them with 100% success without mo Adversarial examples are not more visually than an undefended network.

1 Introduction

It is an open question how to train neural networks so they will be robust to adversarial examples [11]. Recently, three defenses have been proposed to make neural networks robust to adversarial examples:

- MagNet [8] was proposed as an approach to make neural networks robust against adversarial examples through two complementary approaches: adversarial examples near the data manifold are *reformed* to lie on the data manifold that are classified correctly, whereas adversarial examples far away from the data manifold are *detected* and rejected before classification. MagNet does not argue robustness in the white-box setting; rather, the authors argue that MagNet is robust in the grey-box setting where the adversary is aware the defense is in place, knows the parameters of the base classifier, but *not* the parameters of the defense.
- An efficient defense [12] was proposed to make neural networks more robust against adversarial examples by performing Gaussian data augmentation during training, and using the BReLU activation function. The authors do not claim perfect security, but claim this makes attacks visually detectable.
- Adversarial Perturbation Elimination GAN (APE-GAN) [10] is similar to MagNet, only adversarial examples are projected onto the data manifold using a Generative Adversarial Network (GAN) [2]. We did not set out to bypass this defense, but found it to be very similar to MagNet and so we analyze it too.

2 Background

We assume familiarity with neural network versarial examples [11], transferability [6], strong attacks against adversarial examples [1] Net [8]. We briefly review the key details and

Notation Let $F(x) = y$ be a neural networ classification outputting a probability distribu the second-to-last layer (the layer before the th layer) Z, so that $F(x) = \text{softmax}(Z(x))$. Eac corresponds to the predicted probability that x is labelled as class i. Let $C(x) = \arg\max$ correspond to the classification of x on F. In we are concerned with neural networks used images (on MNIST and CIFAR-10).

Adversarial examples [11] are instances x very close to a normal instance x with respec distance metric (L_2 distance, in this paper), $C(x') = t$ for any target t chosen by the adver

We generate adversarial examples with C Wagner's L_2 attack algorithm [1]. Specifically

$$\text{minimize} \quad \|x' - x\|_2^2 + c \cdot \ell(x')$$

- Formulation of adversarial attacks and defenses as optimizing *surrogates* of the true adversarial risk

On the Robustness of the CVPR 2018 White-Box Adversarial Example Defenses

Anish Athalye [*1] Nicholas Carlini [*2]

Abstract

Neural networks are known to be vulnerable to adversarial examples. In this note, we evaluate the two white-box defenses that appeared at CVPR 2018 and find they are ineffective: when applying existing techniques, we can reduce the accuracy of the defended models to 0%.

1. Introduction

Training neural networks so they will be robust to adversarial examples (Szegedy et al., 2013) is a major challenge. Two defenses that appear at CVPR 2018 attempt to address this problem: "Deflecting Adversarial Attacks with Pixel Deflection" (Prakash et al., 2018) and "Defense against Adversarial Attacks Using High-Level Representation Guided Denoiser" (Liao et al., 2018).

In this note, we show these two defenses are not effective in the white-box threat model. We construct adversarial examples that reduce the classifier accuracy to 0% on the ImageNet dataset (Deng et al., 2009) when bounded by a small ℓ_∞ perturbation of $4/255$, a stricter bound than considered in the original papers. Our attacks can construct targeted adversarial examples with over 97% success.

Our methods are a direct application of existing techniques.

2. Background

We assume familiarity with neural networks, adversarial examples (Szegedy et al., 2013), generating strong attacks against adversarial examples (Madry et al., 2018), and computing adversarial examples for neural networks with non-differentiable layers (Athalye et al., 2018). We briefly review the key details and notation.

Adversarial examples (Szegedy et al., 2013) are instances x' that are very close to an instance x with respect to some distance metric (ℓ_∞ distance, in this paper), but where the classification of x' is not the same as the classification of x.

[*]Equal contribution [1]Massachusetts Institute of Technology [2]University of California, Berkeley. Correspondence to: Anish Athalye <aathalye@mit.edu>, Nicholas Carlini <npc@cs.berkeley.edu>.

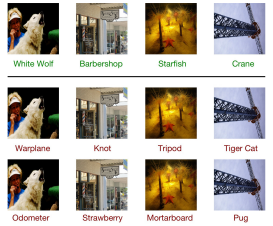


Figure 1. Original images from ImageNet validation set (row 1). Targeted adversarial examples (with randomly chosen targets) for Pixel Deflection (row 2) and High-level representation Guided Denoiser (row 3), with a ℓ_∞ perturbation of $\epsilon = 4/255$.

Targeted adversarial examples are instances x' whose label is equal to a given target label t.

We examine two defenses: Pixel Deflection and High-level Representation Guided Denoiser. We are grateful to the authors of these defenses for releasing their source code and pre-trained models.

Pixel Deflection (Prakash et al., 2018) proposes a non-differentiable preprocessing of inputs. Some pixels (a tunable hyperparameter) are randomly replaced with near-by pixels. This resulting image is often noisy, and to restore accuracy, a denoising operation is applied.

High-level representation Guided Denoiser (HGR) (Liao et al., 2018) proposes denoising inputs using a trained neural network before passing them to a standard classifier. This denoiser is a differentiable, non-randomized neural network. This defense has also been evaluated by Uesato et al. (2018) and found to be ineffective.

2.1. Methods

We evaluate these defenses under the white-box threat model. We generate adversarial examples with Projected

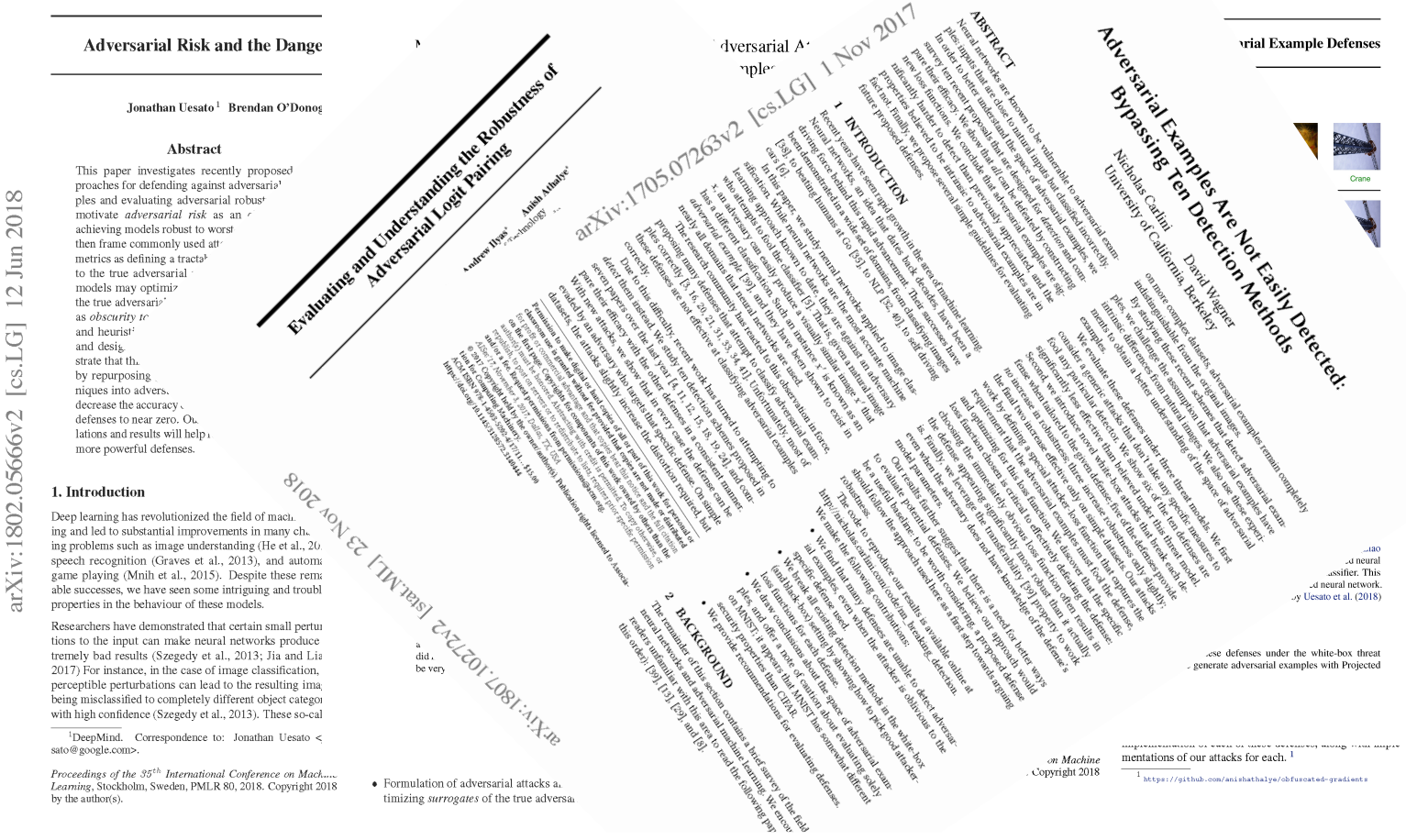
rnational Conference on Machine
1, PMLR 80, 2018.

mentations of our attacks for each. [1]

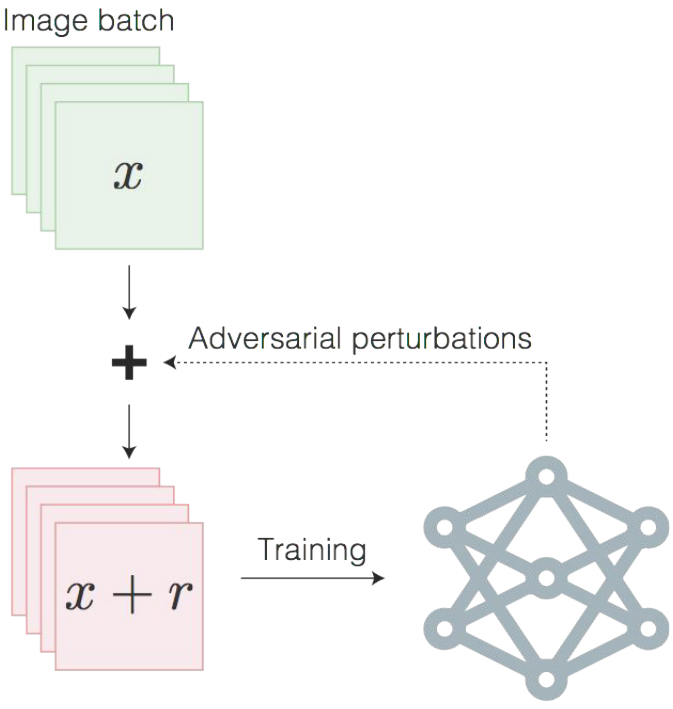
[1] https://github.com/anishathalye/obfuscated-gradients

... but designing defense mechanisms is hard!

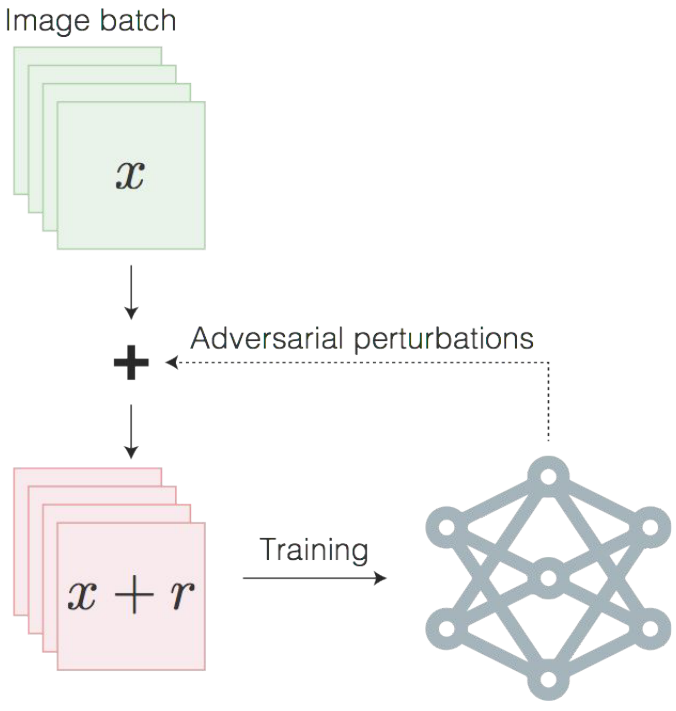
Despite the huge number of proposed defenses, state-of-the-art classifiers are still vulnerable to small perturbations.



Adversarial training



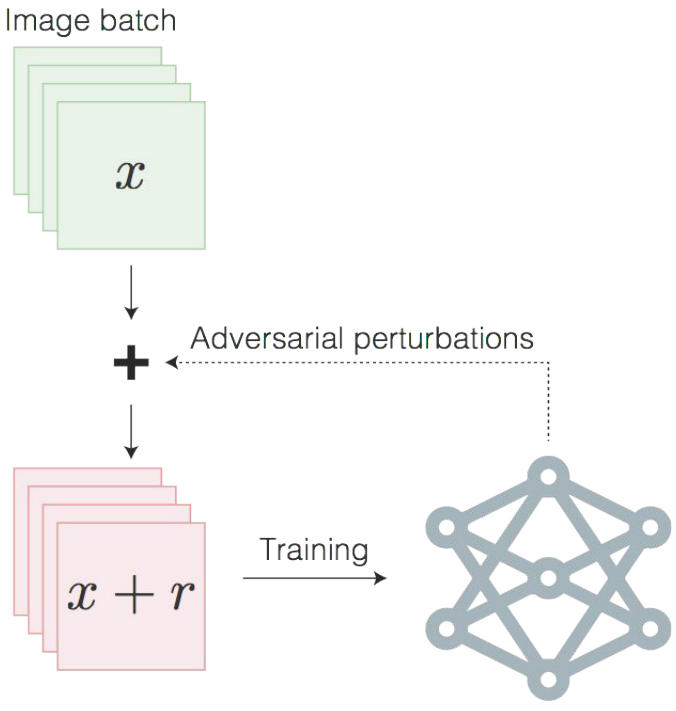
Adversarial training



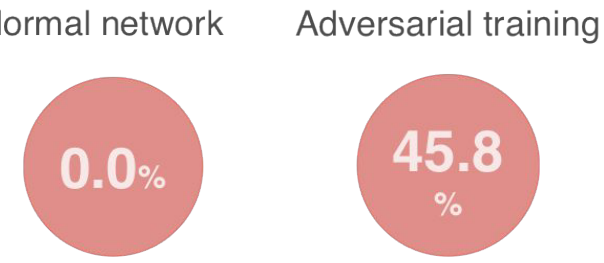
Normal network

Adversarial accuracy (CIFAR-10):

Adversarial training

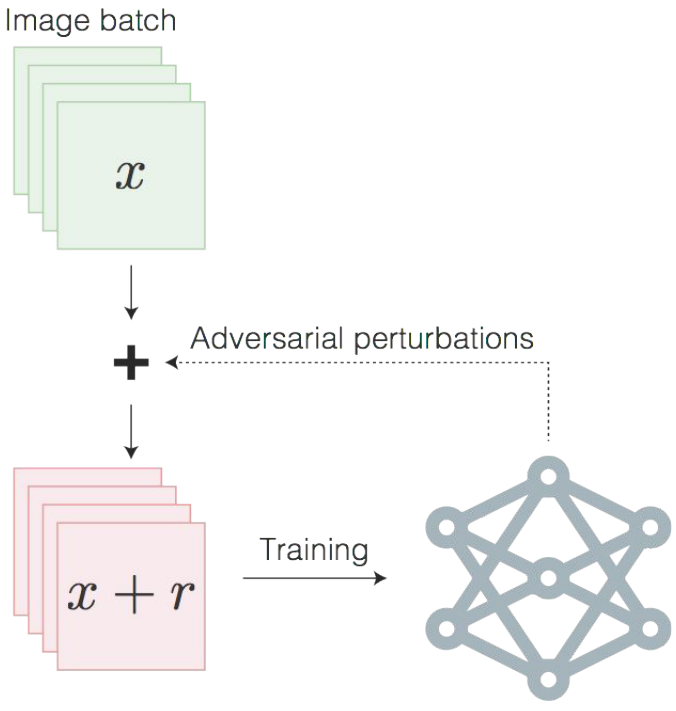


Adversarial accuracy (CIFAR-10):



[Madry et. al., 2017]

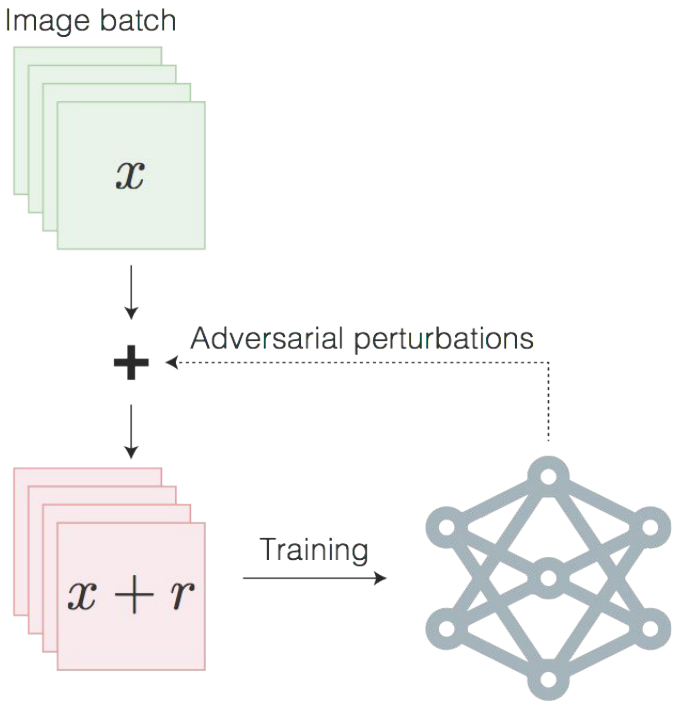
Adversarial training



[Madry et. al., 2017]

Adversarial training leads to state-of-the art robustness to adversarial perturbations.

Adversarial training

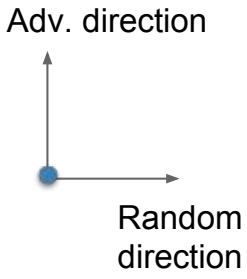


[Madry et. al., 2017]

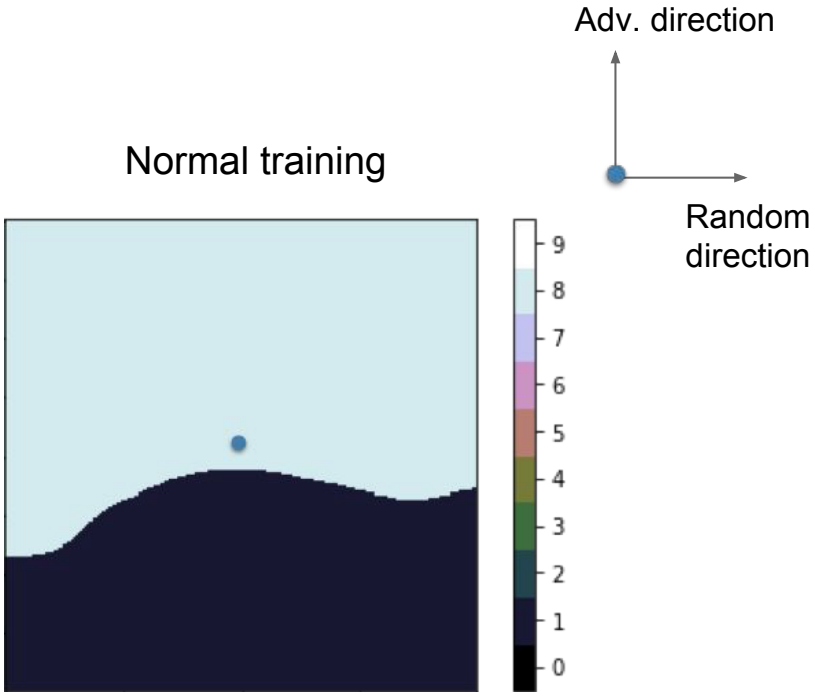
Adversarial training leads to state-of-the art robustness to adversarial perturbations.

But what does it actually do?

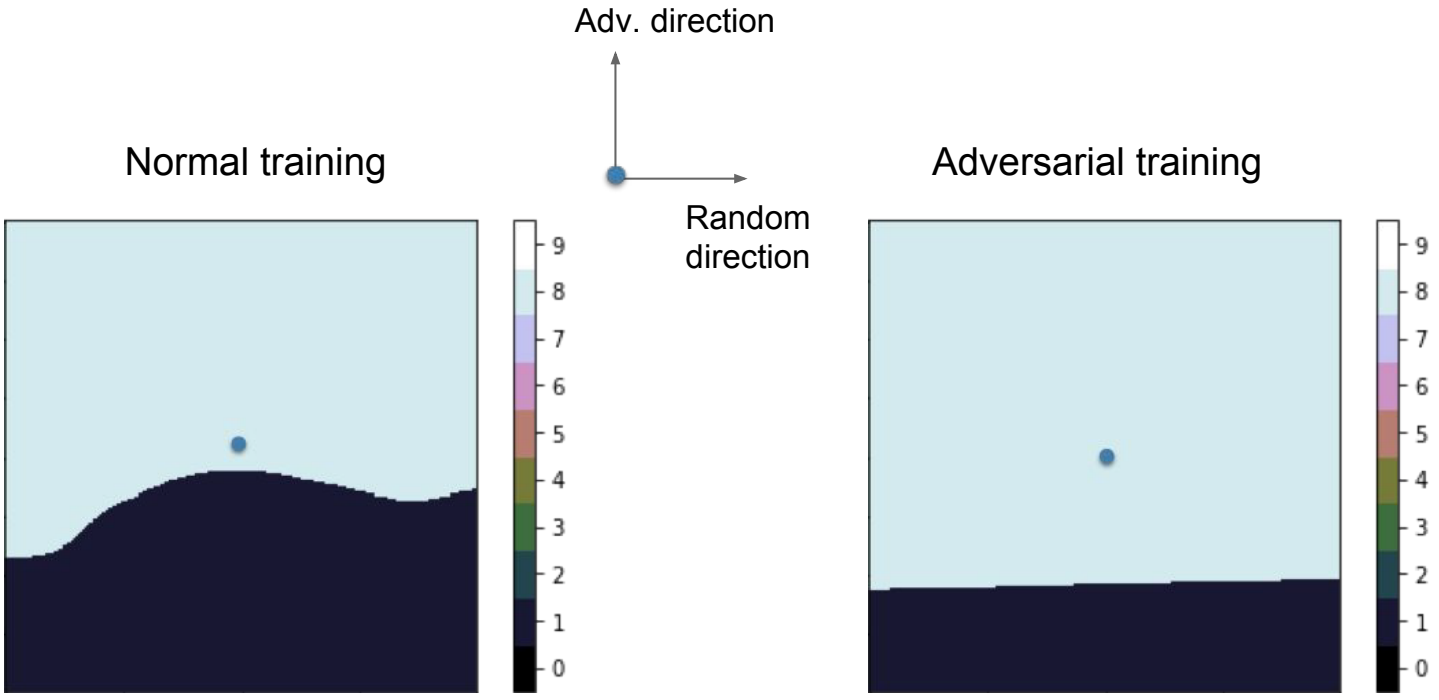
Decision boundaries



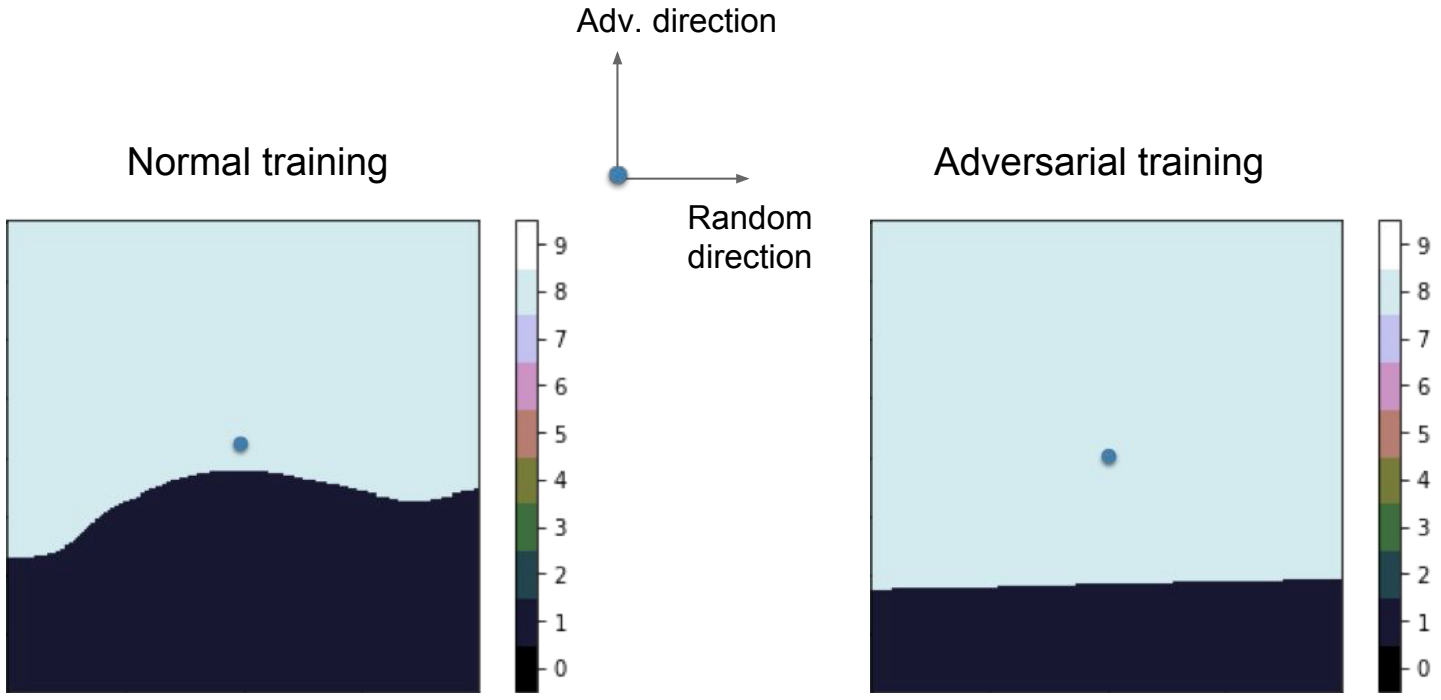
Decision boundaries



Decision boundaries



Decision boundaries



After adversarial training, the decision boundaries are flatter and more regular.

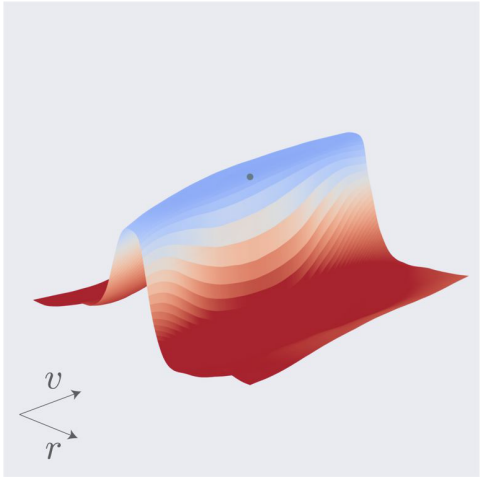
Effect of adversarial training on loss landscape

$$\ell(x) = \mathrm{CE}(\underbrace{f_\theta(x)}_{\text{Logit}}, \underbrace{y}_{\text{Label}})$$

Effect of adversarial training on loss landscape

$$\ell(x) = \mathrm{CE}(\underbrace{f_\theta(x)}_{\text{Logit}}, \underbrace{y}_{\text{Label}})$$

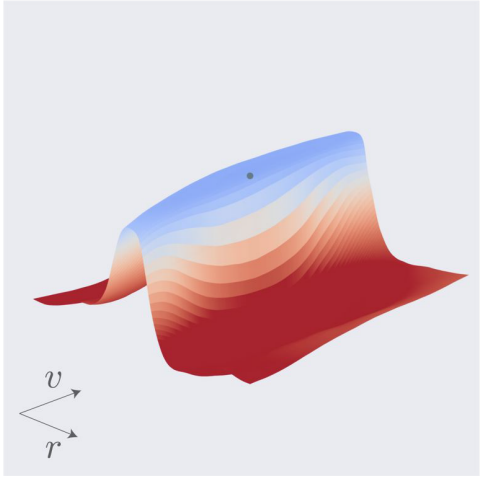
Before adv. fine-tuning



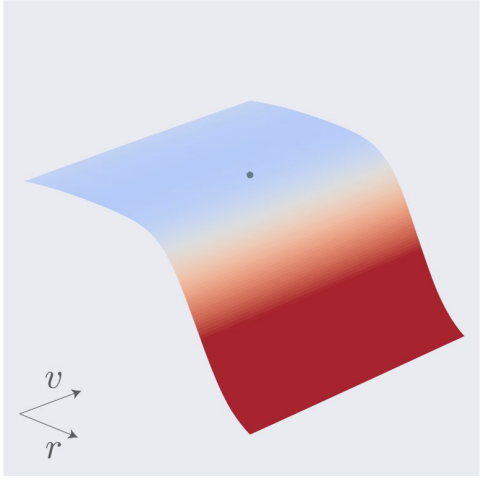
Effect of adversarial training on loss landscape

$$\ell(x) = \mathrm{CE}(\underbrace{f_\theta(x)}_{\text{Logit}}, \underbrace{y}_{\text{Label}})$$

Before adv. fine-tuning



After adv. fine-tuning



Effect of adversarial training on loss landscape (Cont'd)

Quantitative analysis: curvature decrease with adversarial training

We compute the Hessian matrix at a test point x with respect **to inputs**.

$$H = \left(\frac{\partial^2 \ell}{\partial x_i \partial x_j} \right)$$

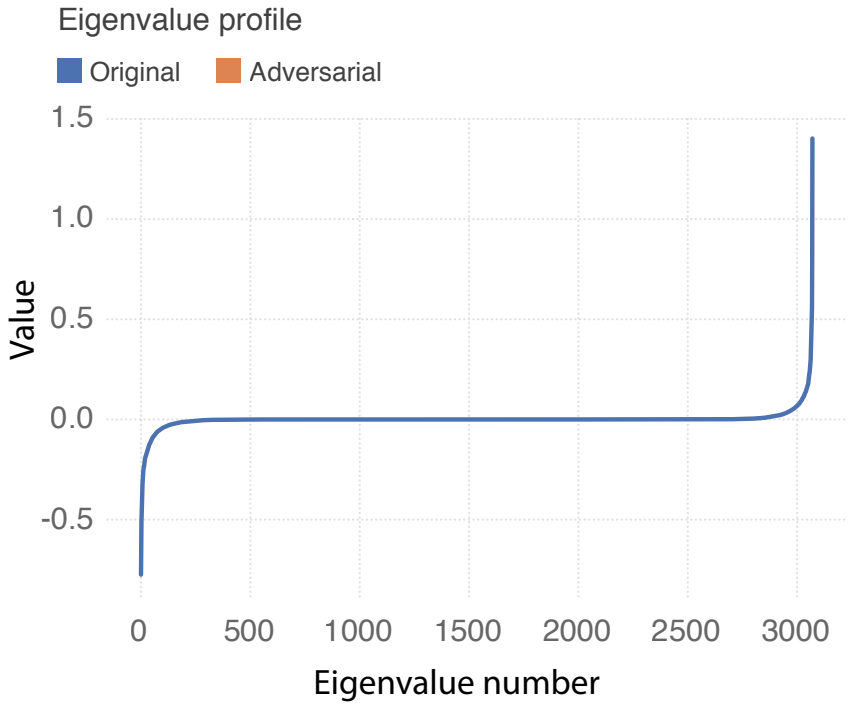
The eigenvalues of H are the curvature of ℓ in the vicinity of x.

Quantitative analysis: curvature decrease with adversarial training

We compute the Hessian matrix at a test point x with respect **to inputs**.

$$H = \left(\frac{\partial^2 \ell}{\partial x_i \partial x_j} \right)$$

The eigenvalues of H are the curvature of ℓ in the vicinity of x.

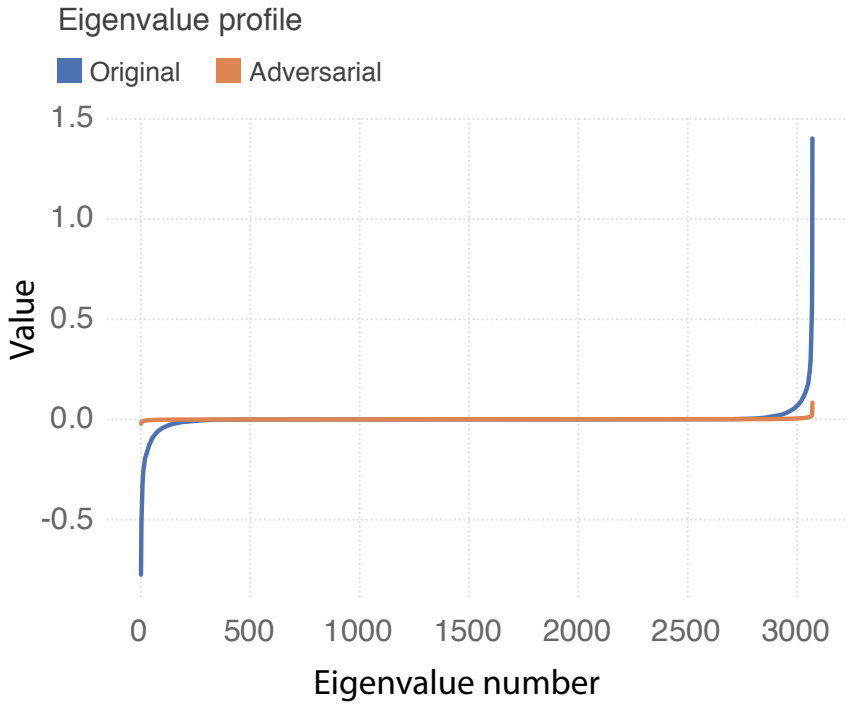


Quantitative analysis: curvature decrease with adversarial training

We compute the Hessian matrix at a test point x with respect **to inputs**.

$$H = \left(\frac{\partial^2 \ell}{\partial x_i \partial x_j} \right)$$

The eigenvalues of H are the curvature of ℓ in the vicinity of x.



Relation between curvature and robustness

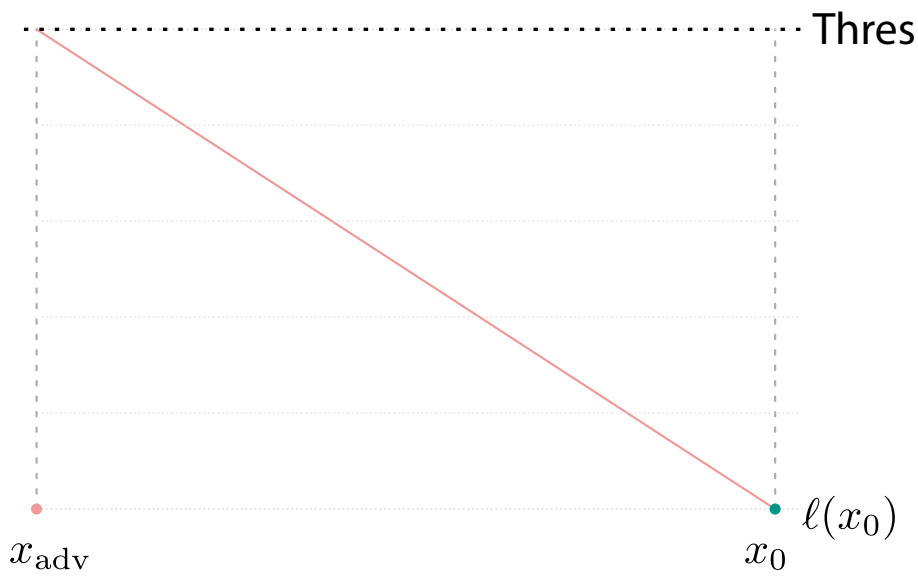
- Locally quadratic approximation of the loss function

Relation between curvature and robustness

- Locally quadratic approximation of the loss function
- We derive upper and lower bounds on the minimal perturbation required to fool a classifier.

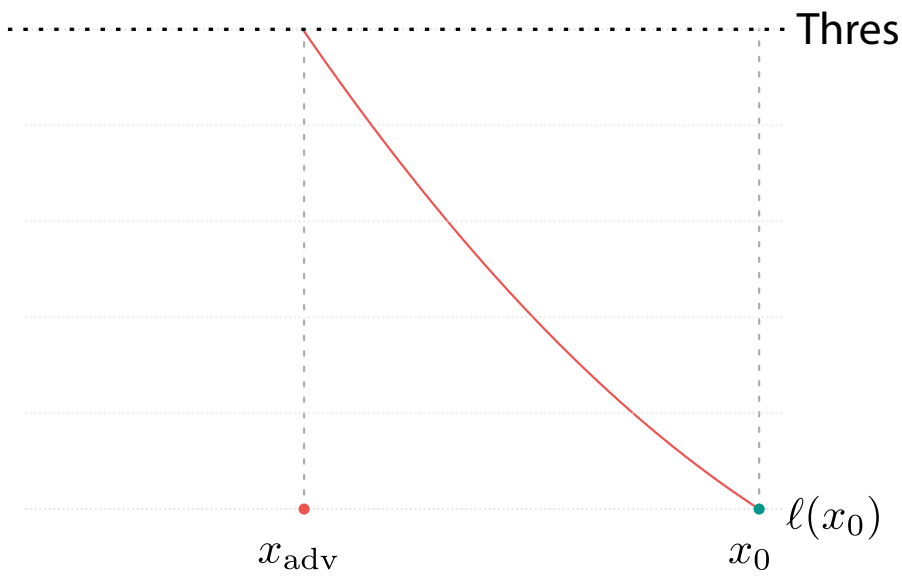
Relation between curvature and robustness

- Locally quadratic approximation of the loss function
- We derive upper and lower bounds on the minimal perturbation required to fool a classifier.



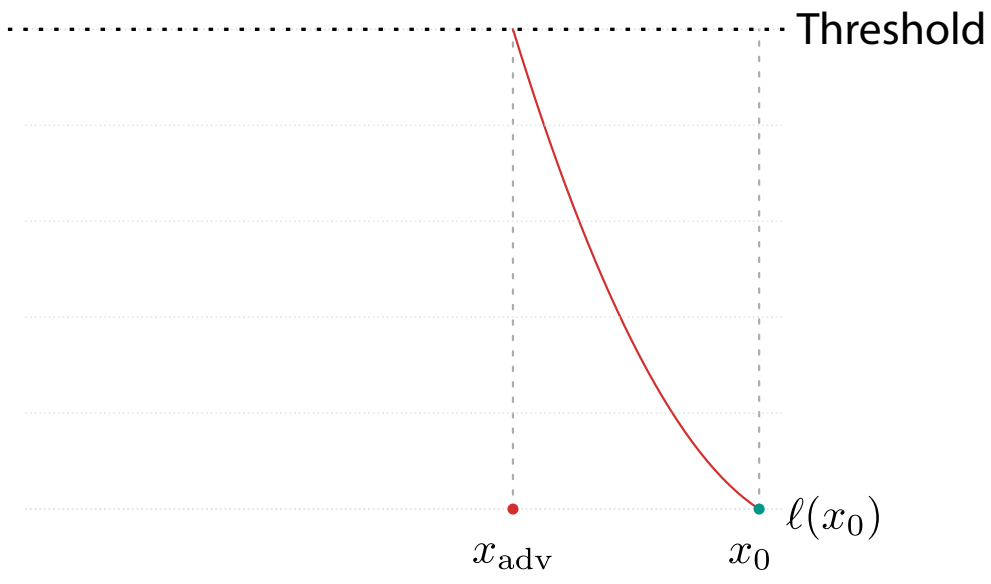
Relation between curvature and robustness

- Locally quadratic approximation of the loss function
- We derive upper and lower bounds on the minimal perturbation required to fool a classifier.



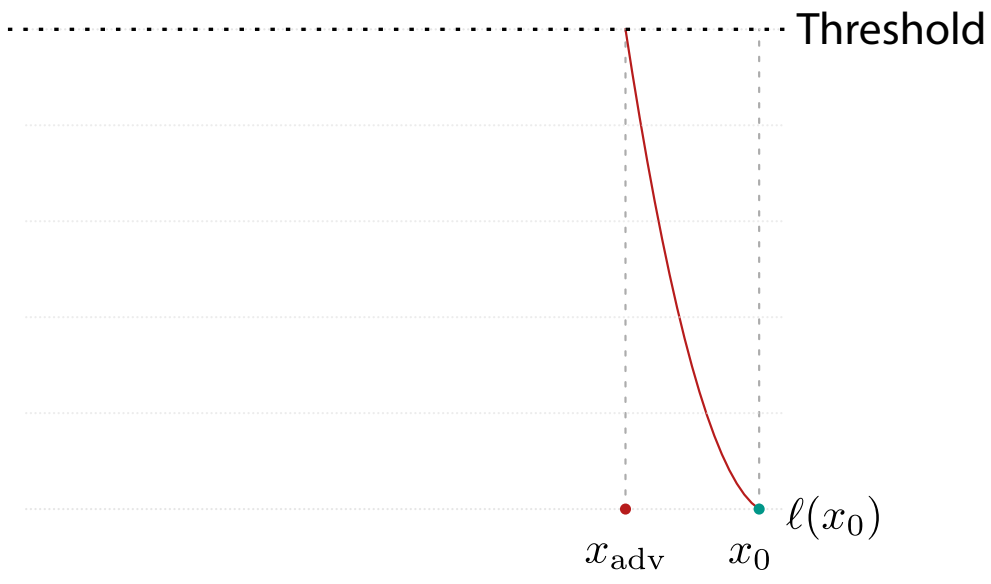
Relation between curvature and robustness

- Locally quadratic approximation of the loss function
- We derive upper and lower bounds on the minimal perturbation required to fool a classifier.



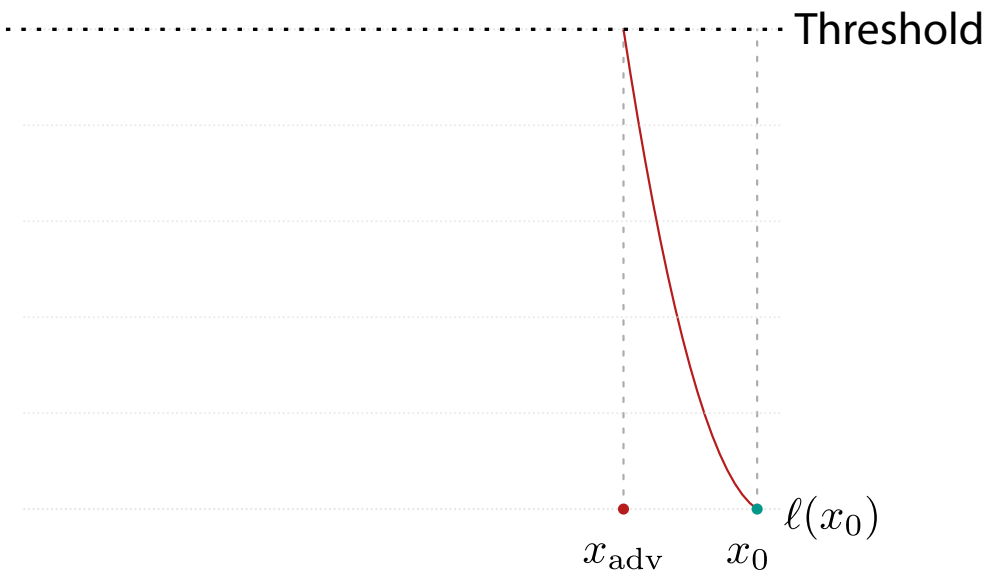
Relation between curvature and robustness

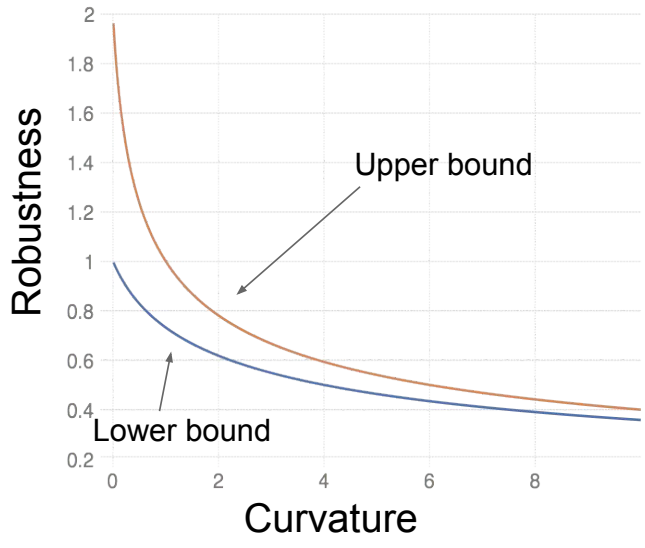
- Locally quadratic approximation of the loss function
- We derive upper and lower bounds on the minimal perturbation required to fool a classifier.



Relation between curvature and robustness

- Locally quadratic approximation of the loss function
- We derive upper and lower bounds on the minimal perturbation required to fool a classifier.





How important is the curvature decrease?

Is the curvature decrease **the main effect of adversarial training** leading to improved robustness?

How important is the curvature decrease?

Is the curvature decrease **the main effect of adversarial training** leading to improved robustness?
$\rightarrow$ we regularize explicitly for the curvature.

How important is the curvature decrease?

Is the curvature decrease **the main effect of adversarial training** leading to improved robustness?
$\rightarrow$ we regularize explicitly for the curvature.

- Idea: Regularize the norm of the Hessian of the loss wrt inputs.

How important is the curvature decrease?

Is the curvature decrease **the main effect of adversarial training** leading to improved robustness?
$\rightarrow$ we regularize explicitly for the curvature.

- Idea: Regularize the norm of the Hessian of the loss wrt inputs.
- Use Hutichinson's estimator

$$\|H\|_F = \sqrt{\mathop{\mathbb{E}}_{z \sim \mathcal{N}(0,I)} \|Hz\|_2^2}$$

How important is the curvature decrease?

Is the curvature decrease **the main effect of adversarial training** leading to improved robustness?
$\rightarrow$ we regularize explicitly for the curvature.

- Idea: Regularize the norm of the Hessian of the loss wrt inputs.
- Use Hutichinson's estimator

$$\|H\|_F = \sqrt{\mathop{\mathbb{E}}_{z \sim \mathcal{N}(0,I)} \|Hz\|_2^2}$$

- In practice:
  - Compute Hessian-vector products with finite difference
  - Selective sampling on directions corresponding to high curvature

How important is the curvature decrease?

Is the curvature decrease **the main effect of adversarial training** leading to improved robustness?
$\rightarrow$ we regularize explicitly for the curvature.

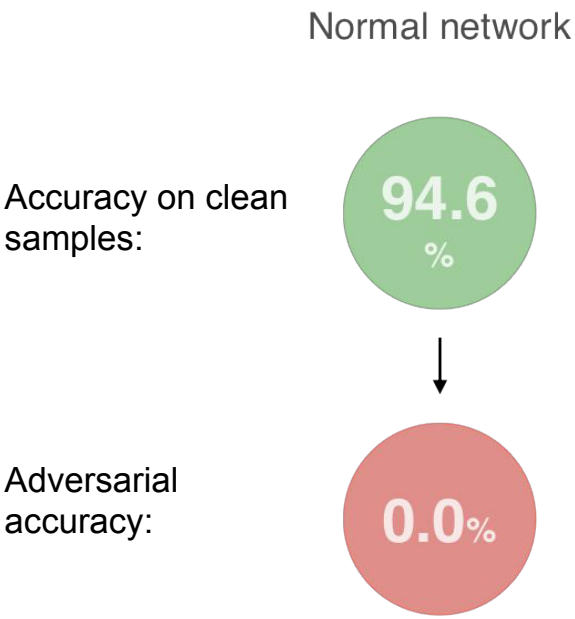
- Idea: Regularize the norm of the Hessian of the loss wrt inputs.
- Use Hutichinson's estimator

$$\|H\|_F = \sqrt{\mathop{\mathbb{E}}_{z \sim \mathcal{N}(0,I)} \|Hz\|_2^2}$$

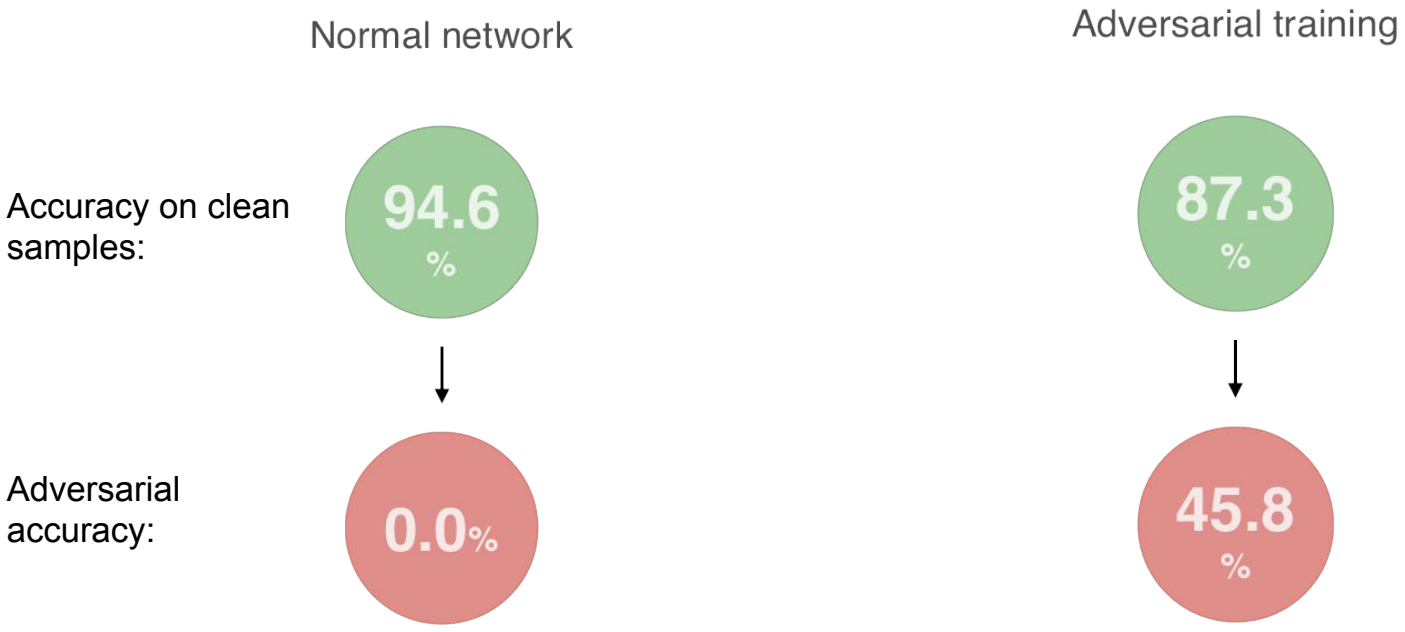
- In practice:
  - Compute Hessian-vector products with finite difference
  - Selective sampling on directions corresponding to high curvature

CURE: Regularize using $\ell_r = \|\nabla \ell(x + hz) - \nabla \ell(x)\|$

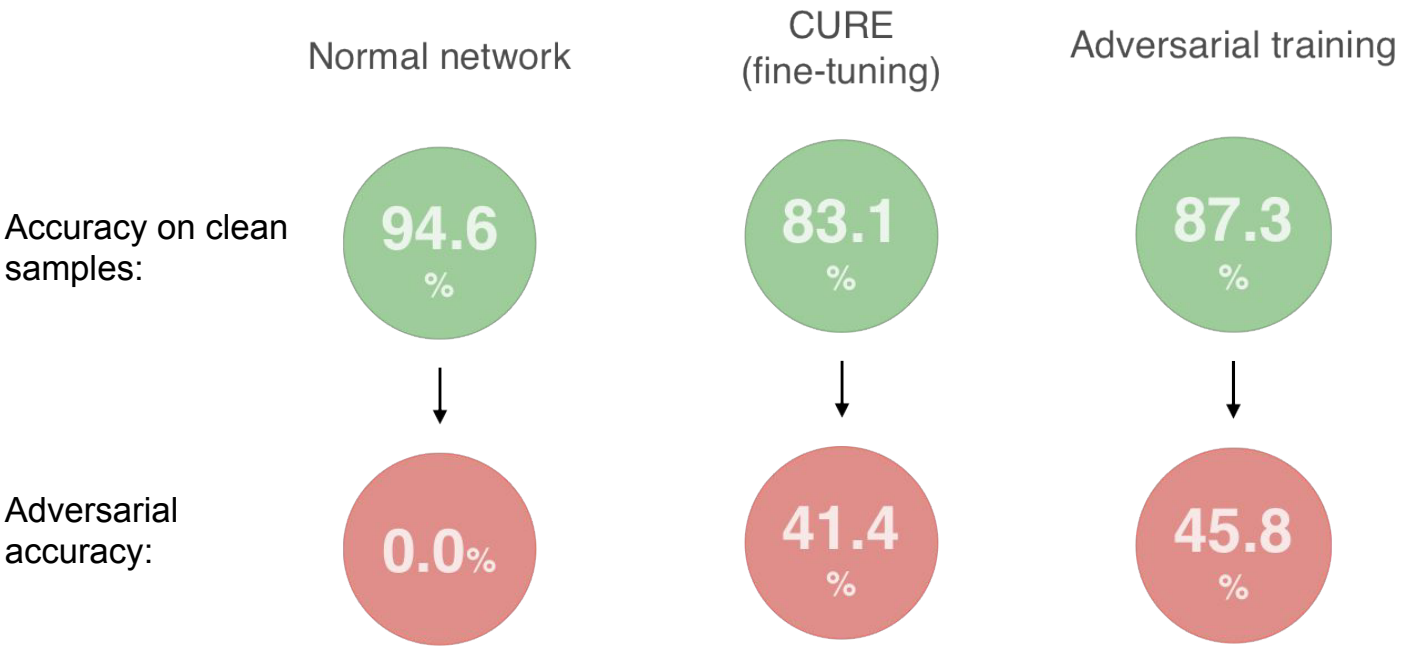
CUREing deep networks trained on CIFAR-10



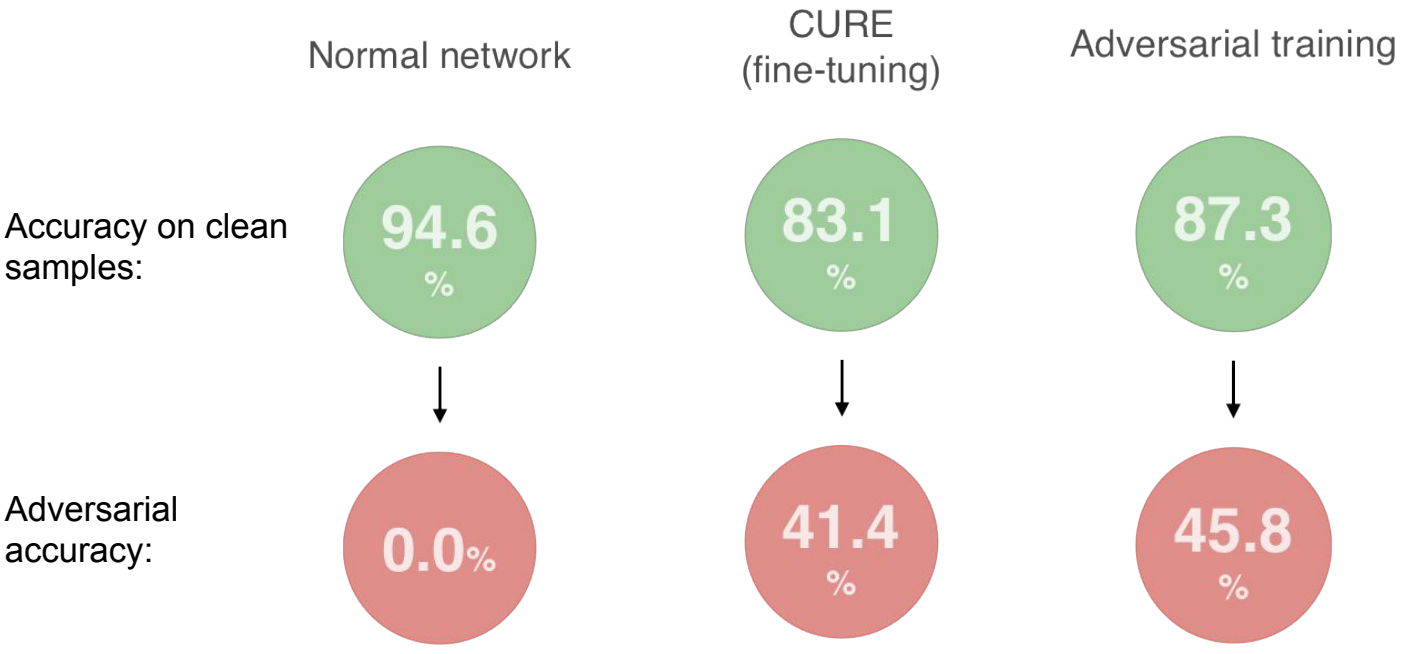
CUREing deep networks trained on CIFAR-10

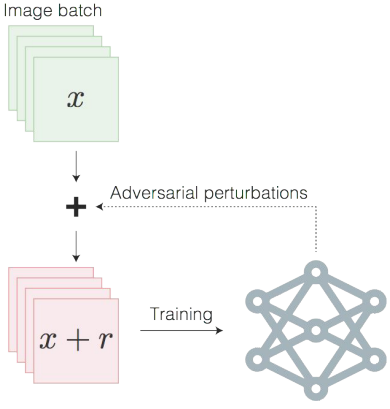


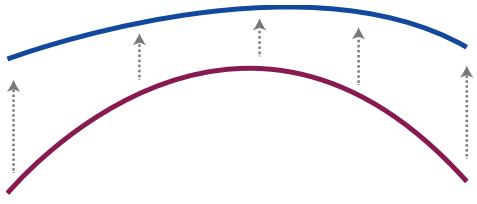
CUREing deep networks trained on CIFAR-10



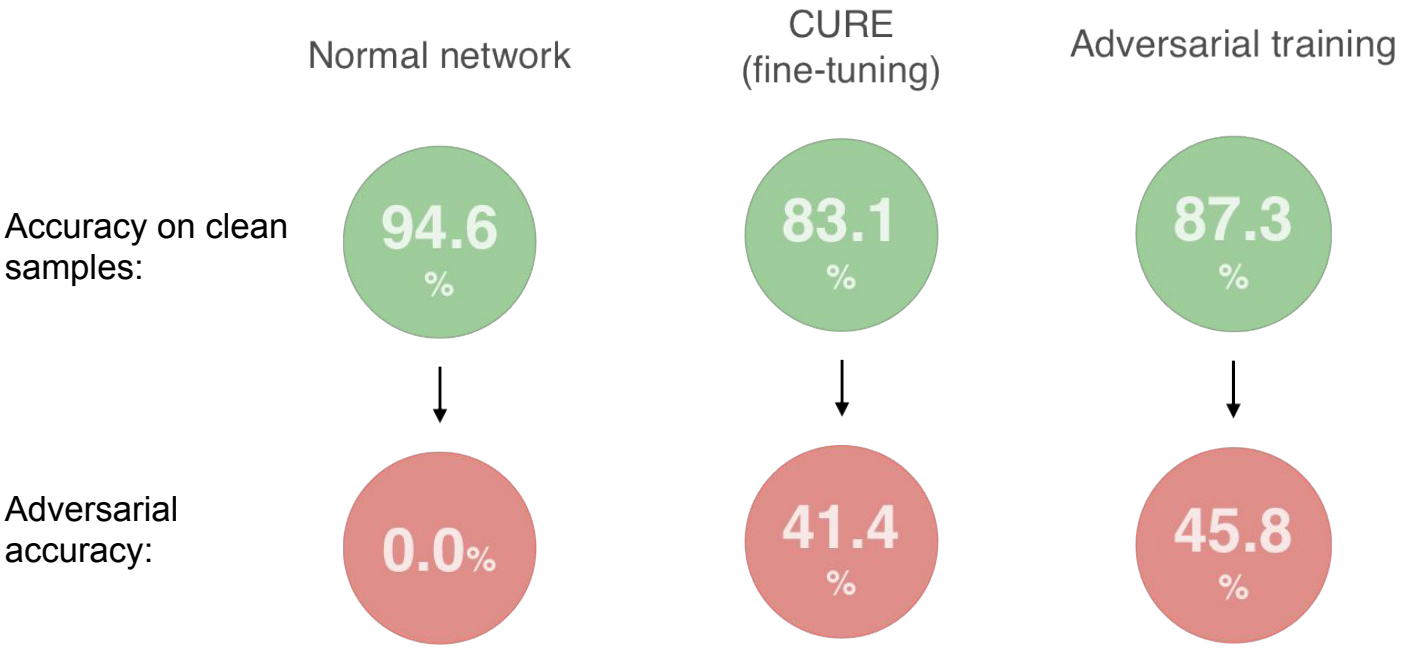
CUREing deep networks trained on CIFAR-10

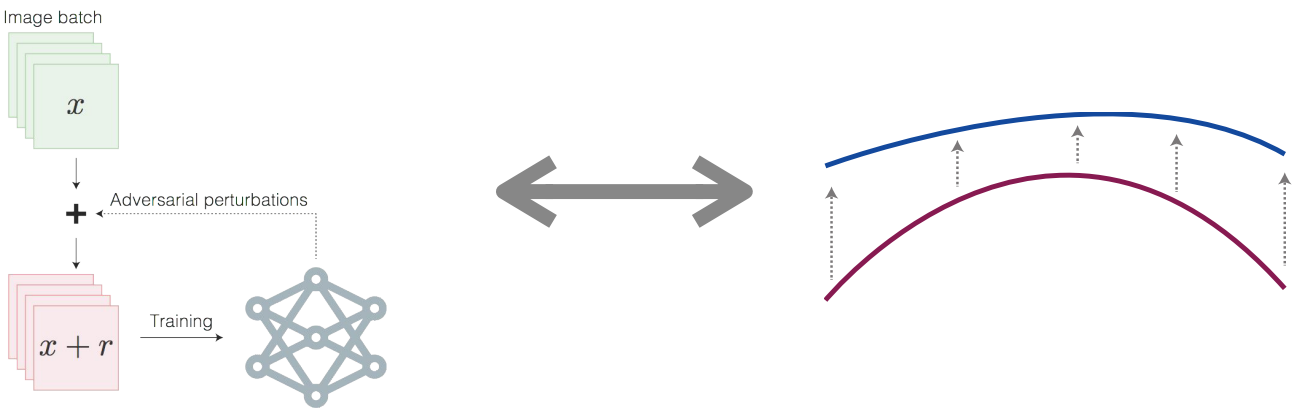






CUREing deep networks trained on CIFAR-10





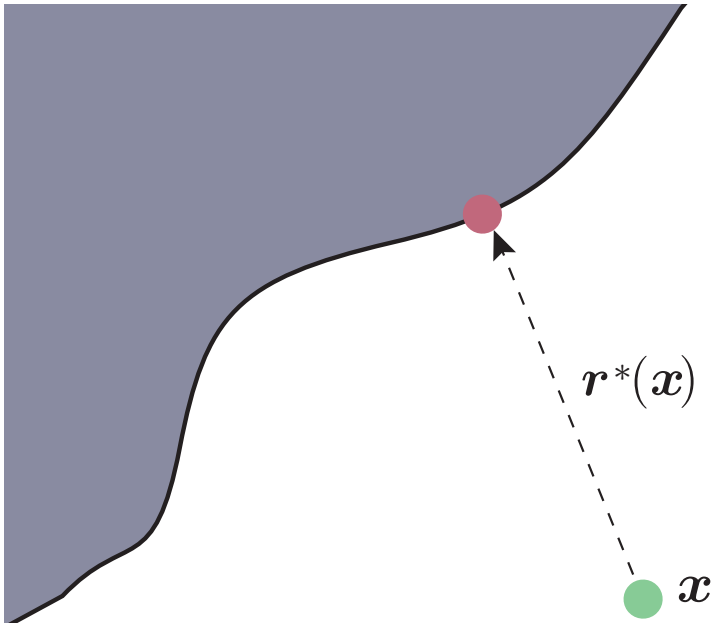
[Moosavi-Dezfooli, et. al., *Robustness via curvature regularization, and vice-versa*, CVPR 2019]

Upper limits on adversarial robustness

- Goal: examine the existence of upper bounds on the robustness to adversarial perturbations.
- Relate to quantities that we better understand/we can better measure: *robustness to random noise*.
- Comparison to the robustness to random noise quantifies the power of an adversary having access to the model vs. no clue about the model.

From adversarial to random noise

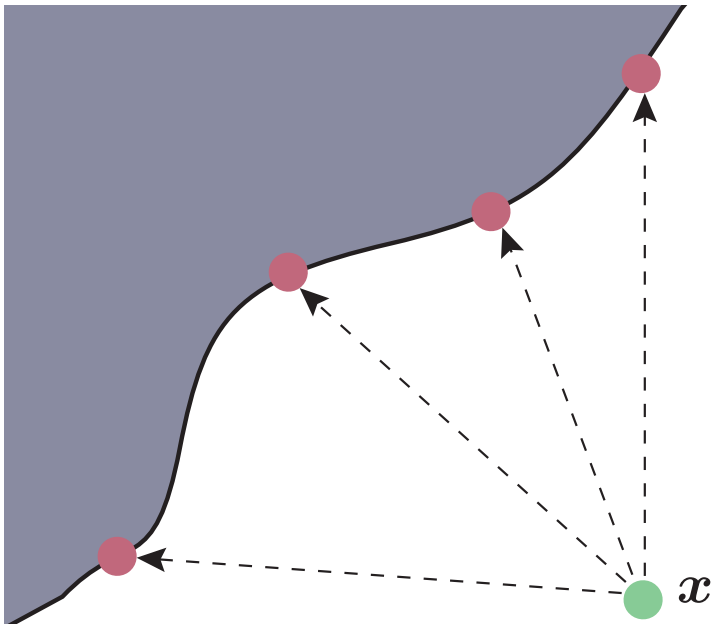
Robustness to adversarial noise



$$\min_r \|r\|_2 \text{ subject to } f(x+r) \neq f(x).$$

From adversarial to random noise

Robustness to random noise



$$\min_t |t| \text{ subject to } f(x + tv) \neq f(x)$$

v uniformly sampled from $\mathbb{S}^{D-1}$.

Linear classifiers

Theorem (Fawzi et. al., NIPS '16, Franceschi et. al., AISTATS '18)

For affine classifiers, we have

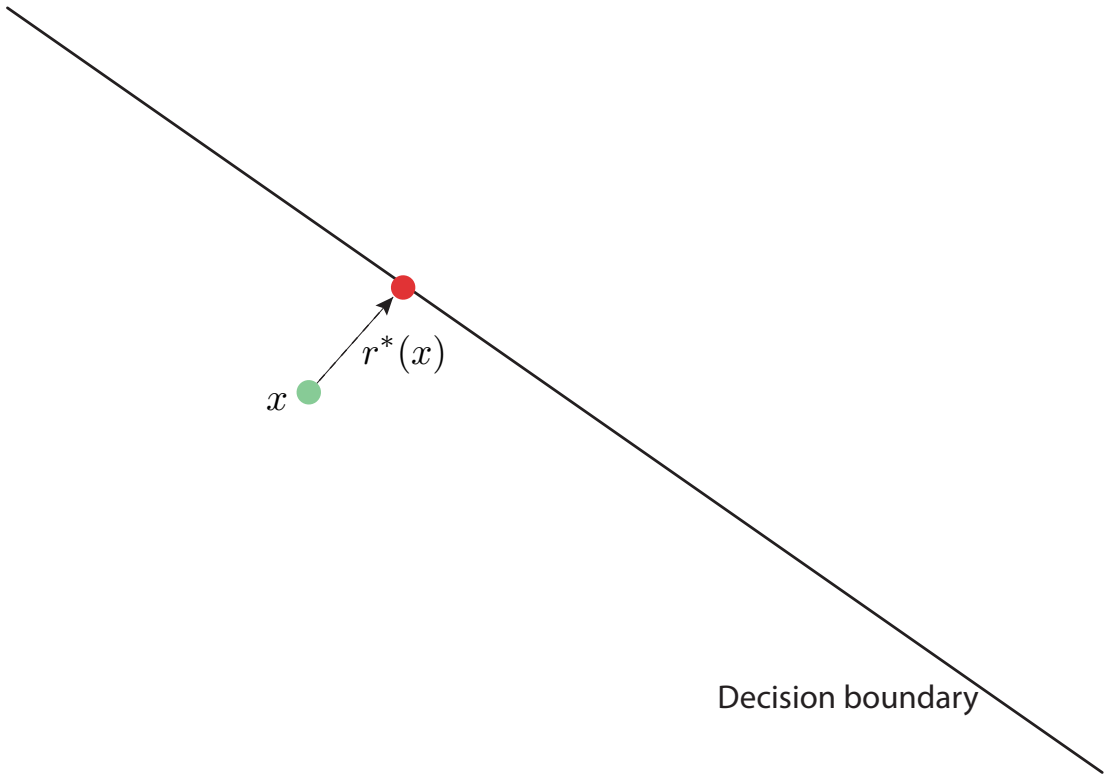
$$\|r^*\|_2 = \Theta\left(\frac{1}{\sqrt{D}}\|r^*_{rand}\|_2\right),$$

with high probability (over the choice of random perturbation).

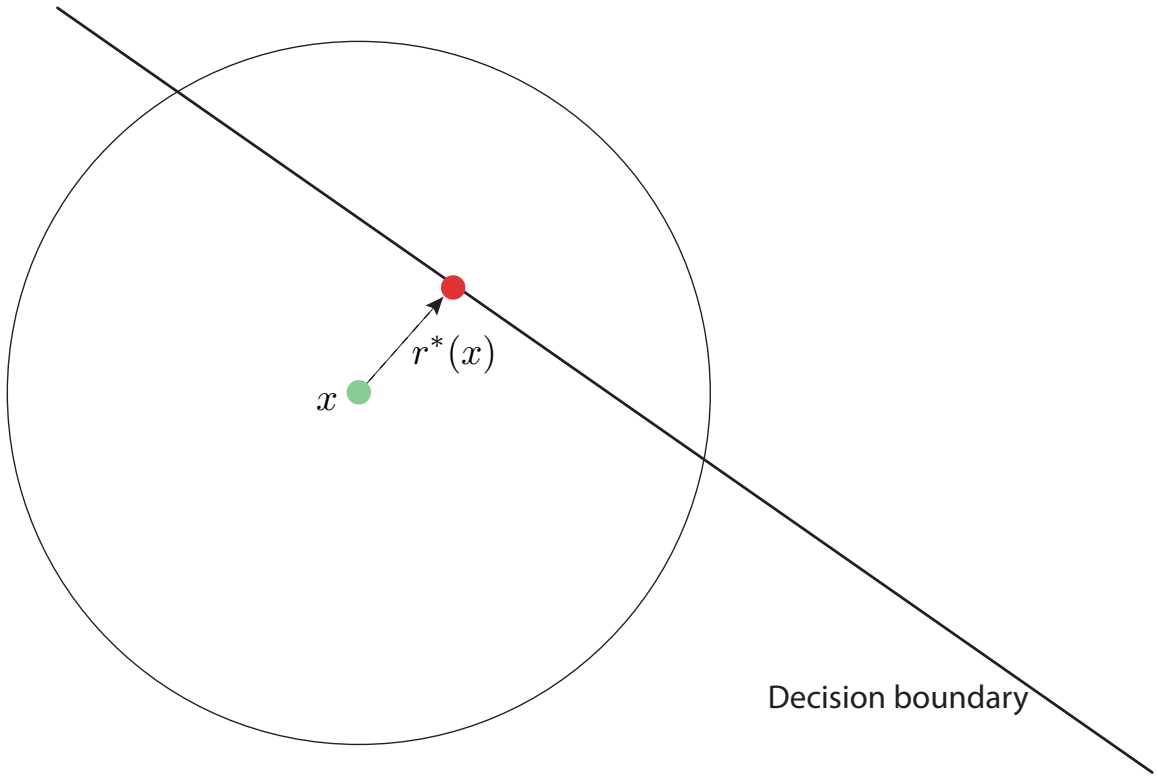
In high dimensions, very large gap between both robustness measures.

When data is bounded, we therefore typically get $\|r^*\|_2 = O\left(\frac{1}{\sqrt{D}}\right) \rightarrow$ Achieving robustness in high dimensions is difficult!

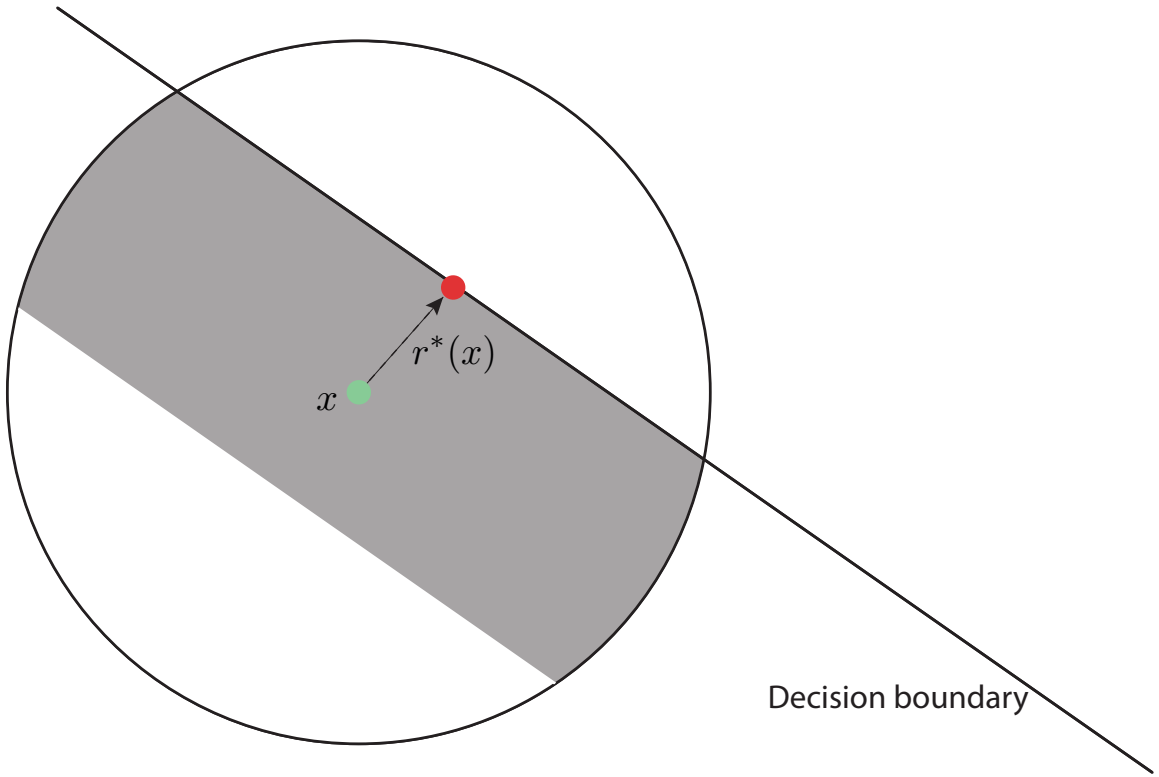
Intuition



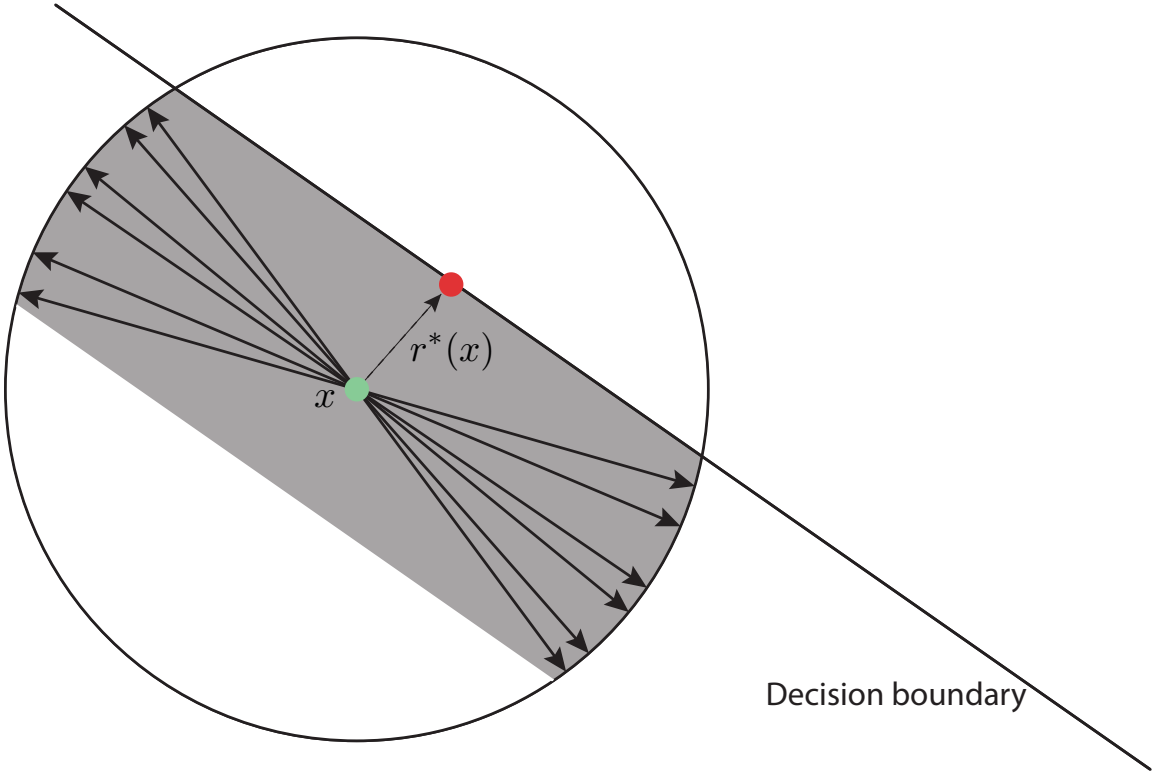
Intuition



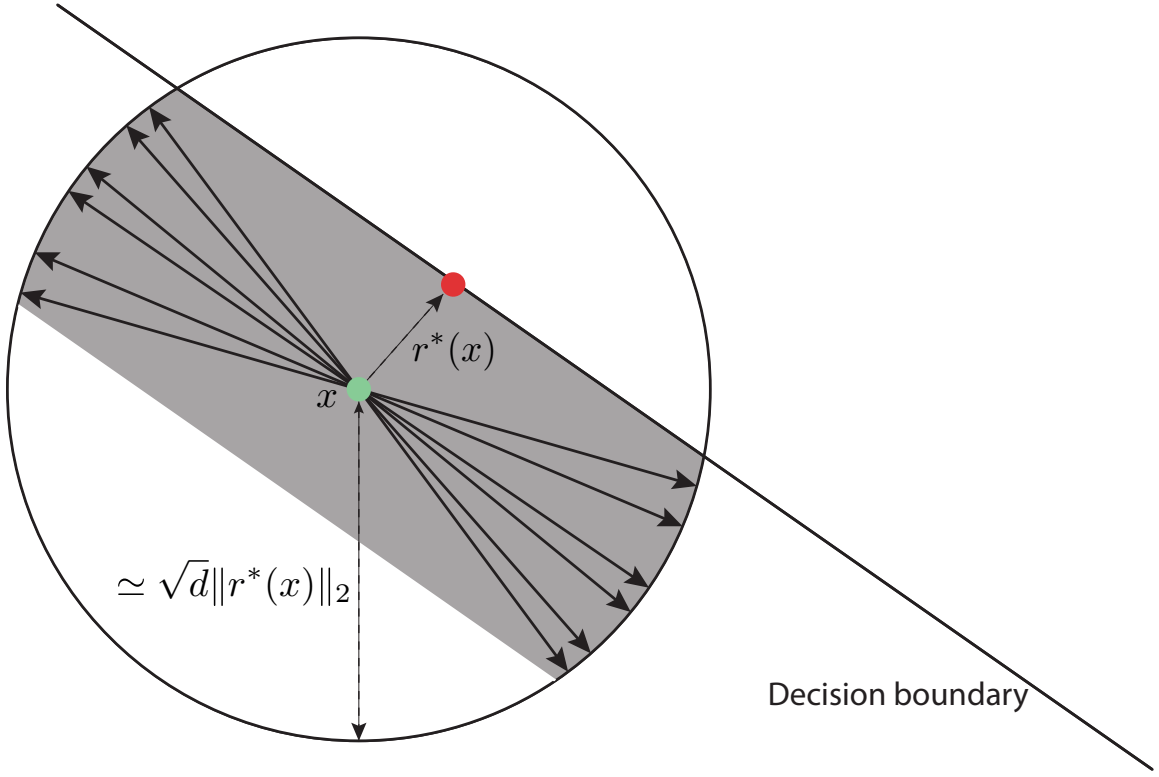
Intuition



Intuition



Intuition



Robustness and geometry of deep networks: SPM'17

DEEP LEARNING FOR VISUAL UNDERSTANDING

Alhussein Fawzi, Seyed-Mohsen Moosavi-Dezfooli,
and Pascal Frossard

The Robustness of Deep Networks

A geometrical perspective

Deep neural networks have recently shown impressive classification performance on a diverse set of visual tasks. When deployed in real-world (noise-prone) environments, it is equally important that these classifiers satisfy robustness guarantees: small perturbations applied to the samples should not yield significant loss to the performance of the predictor. The goal of this article is to discuss the robustness of deep networks to a diverse set of perturbations that may affect the samples in practice, including adversarial perturbations, random noise, and geometric transformations. This article further discusses the recent works that build on the robustness analysis to provide geometric insights on the classifier's decision surface, which help in developing a better understanding of deep networks. Finally, we present recent solutions that attempt to increase the robustness of deep networks. We hope this review article will contribute to shed ding light on the open research challenges in the robustness of deep networks and stir interest in the analysis of their fundamental properties.

Introduction

With the dramatic increase of digital data and the development of new computing architectures, deep learning has been developing rapidly as a predominant framework for data representation that can contribute in solving very diverse tasks. Despite this success, several fundamental properties of deep neural networks are still not understood and have been the subject of intense analysis in recent years. In particular, the robustness of deep networks to various forms of perturbations has

Conclusions

- Very simple strategies to fool classifiers: adversarial, geometric, universal perturbations, ...
- Difficulty of defending against adversarial perturbations; most "defenses" make the classifier robust against specific directions.
- Adversarial training is essentially reducing the curvature of the loss function, leading to an increased robustness.

References

- Szegedy et. al., *Intriguing properties of neural networks*, ICLR 2014
- Fawzi et. al, *Manitest: Are classifiers really invariant?*, BMVC 2015
- Moosavi-Dezfooli et. al., *Deepfool: a simple and accurate approach to fool deep neural networks*, CVPR 2016
- **Fawzi et. al., Robustness of classifiers: from adversarial to random noise, NIPS 2016**
- Moosavi-Dezfooli et. al., *Universal adversarial perturbations*, CVPR 2017
- Uesato et al., *Adversarial risk and the dangers of evaluating against weak attacks*, arXiv 2018
- Fawzi et. al., Adversarial vulnerability of any classifier, NeurIPS 2018
- **Moosavi-Dezfooli et. al., Robustness via curvature regularization, and vice-versa, CVPR 2019**
- **Survey:** [Fawzi et. al., *Robustness of deep networks: a geometric perspective*, IEEE Signal Processing Magazine 2017]