Adversarial examples

The Robust Features Model

Useless
directions

Robust features
Correlated with label even when perturbed

Non-robust features
Correlated with label, but can be flipped via perturbation

Just bugs, too

- **World 1: Adversarial examples exploit directions irrelevant for classification (“bugs”)**

  Adversarial examples occur because classifiers behave poorly off-distribution. They would occur in arbitrary directions, having nothing to do with the true data distribution.

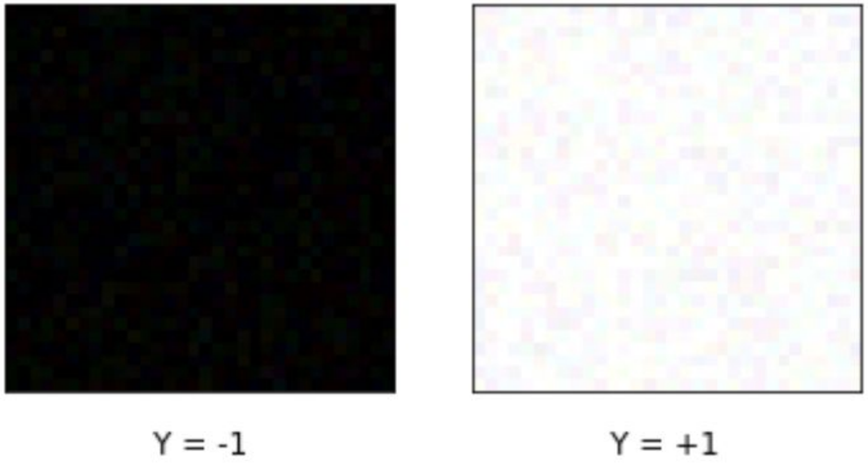
- **World 2: Adversarial examples exploit useful directions for classification (“features”)**

  Adversarial examples occur in directions that are still “on-distribution”, and which contain features of the target class. Perturbation is not purely random. Moreover, we expect that this perturbation transfers to other classifiers trained to distinguish cats vs. dogs.

Paper: P. Nakkiran, Adversarial Examples are Just Bugs, Too, 2019

Adversarial Examples from Robust Features

The problem is to distinguish between CIFAR-sized images that are either all-black or all-white, with a small amount of random pixel noise and label noise.



A sample of images from the distribution.

Adversarial Examples from Robust Features

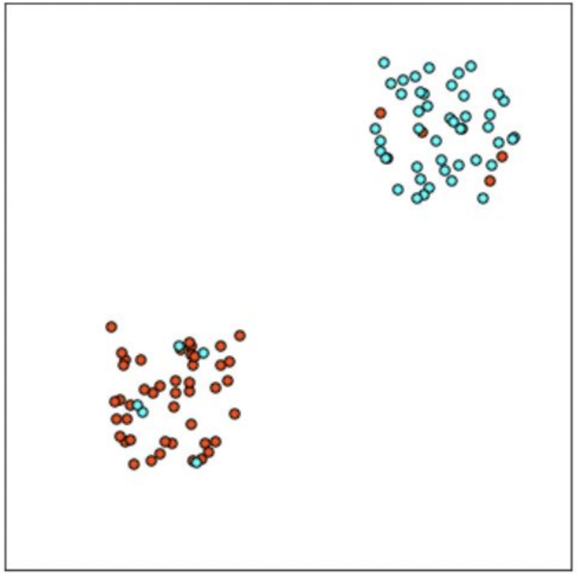
Formally, let the distribution be as follows.

Pick label $Y \in \{\pm 1\}$ uniformly, and let

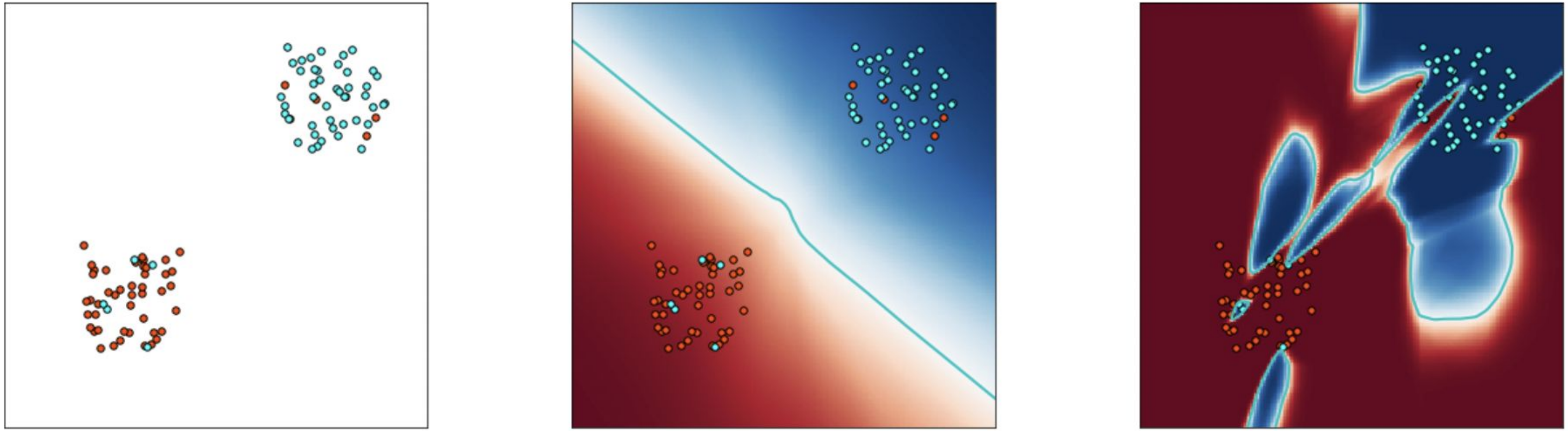
$$X := \begin{cases} (+\vec{1} + \vec{\eta}_{\varepsilon}) \cdot \eta & \text{if } Y = 1 \\ (-\vec{1} + \vec{\eta}_{\varepsilon}) \cdot \eta & \text{if } Y = -1 \end{cases}$$

where $\vec{\eta}_{\varepsilon} \sim [-0.1, +0.1]^d$ is uniform L_∞ pixel noise, and

$\eta \in \{\pm 1\} \sim Bernoulli(0.1)$ is the 10% label noise.



Adversarial Examples from Robust Features



Left: The training set (labels color-coded). Middle: The classifier after 10 SGD steps. Right: The classifier at the end of training. Note that it is overfit, and not robust.

Discontinuities

Image

- $x \in \mathbb{R}^m$

Network

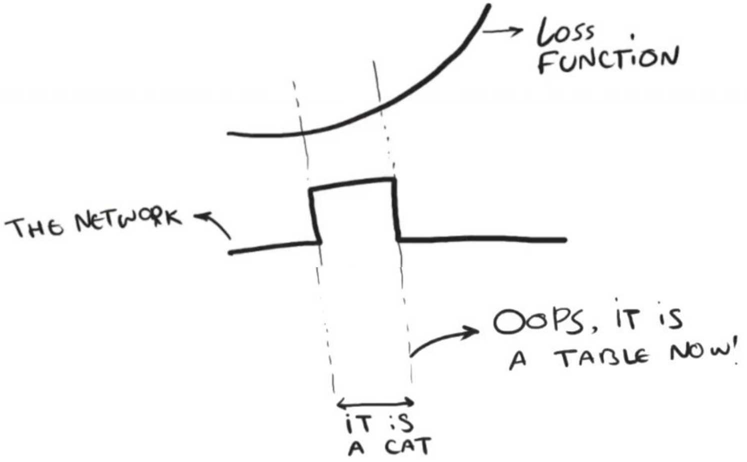
- $f : \mathbb{R}^m \to \{1 ... k\}$

Loss

- $\text{loss}_f : \mathbb{R}^m \times \{1 ... k\} \to \mathbb{R}^+$

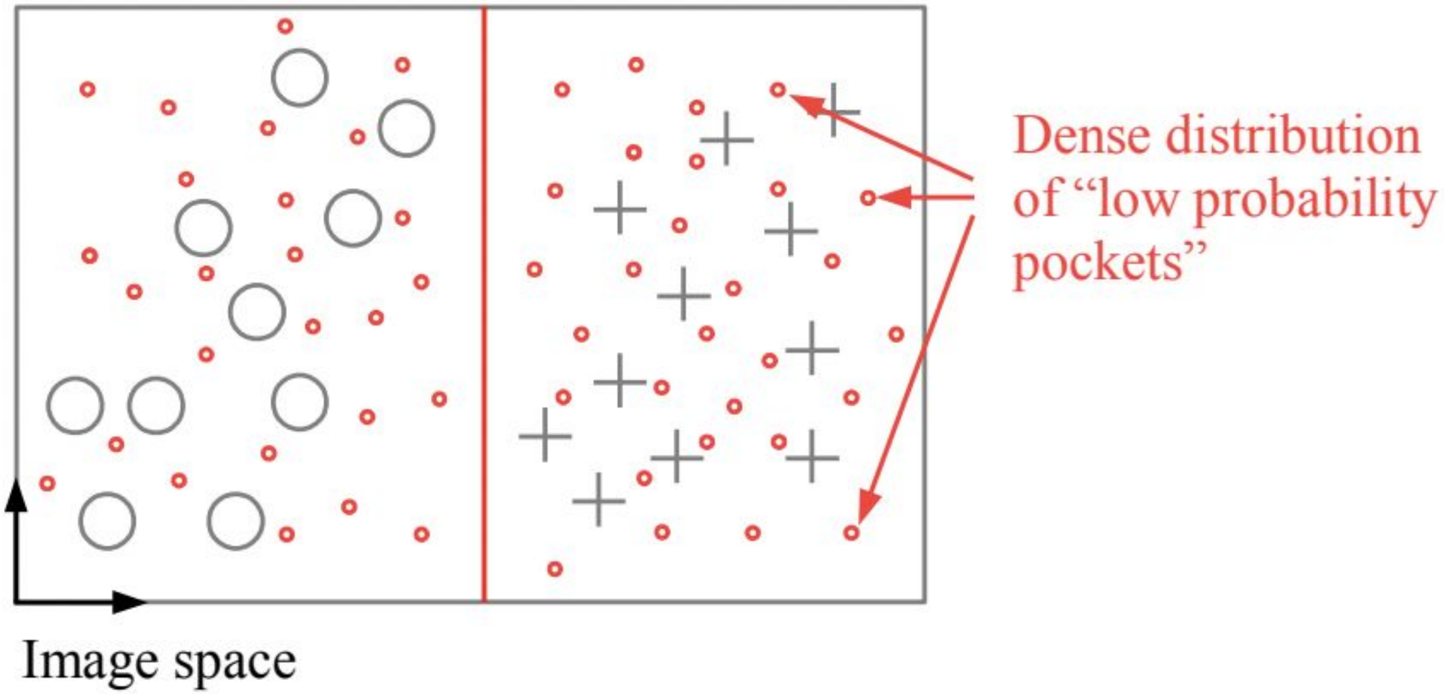
Perturbation $r \in \mathbb{R}^m$ and minimized $\|r\|_2$ such that:

1. $f(x + r) = l$ and $l \neq f(x)$
2. $x + r \in [0, 1]^m$

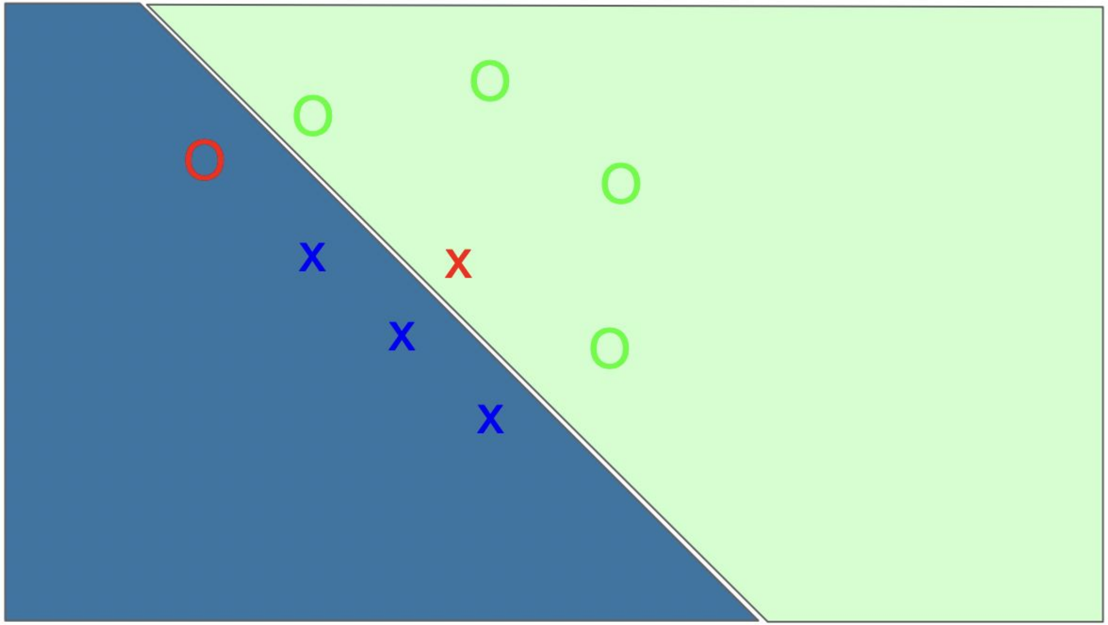


Paper: C. Szegedy, Intriguing properties of neural networks, 2014

Discontinuities



Adversarial Examples from Excessive Linearity

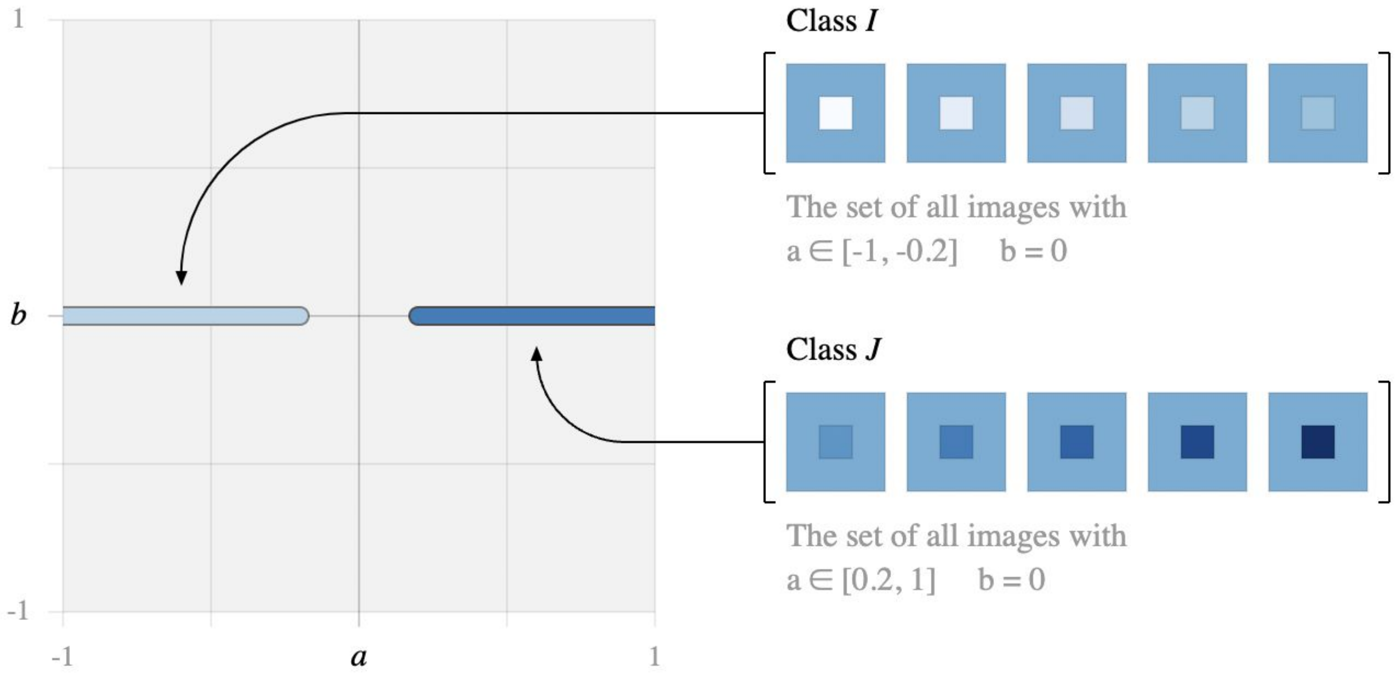


Paper: C. Szegedy, Intriguing properties of neural networks, 2014

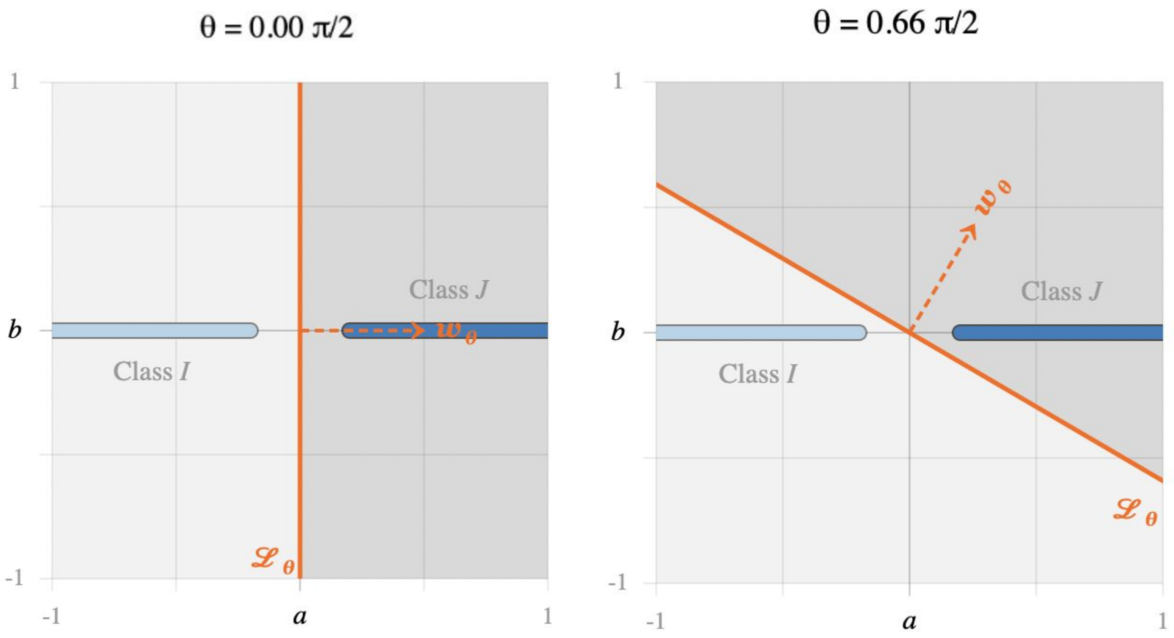
Excessive Linearity

- Let adversarial input x' $= x + \eta$ for some input x.
- For a classifier **F**, we expect **F**(x) = **F**(x') if $||\eta||_\infty < \varepsilon$, for ε small enough to be discarded by the sensor or data storage.
- Dot product of weight w and an adversarial example x' is $w^\mathsf{T}x + w^\mathsf{T}\varepsilon$ (i.e., activation grows by $w^\mathsf{T}\varepsilon$).
  - Put another way, activation grows by εmn, where n is the dimensionality of w, and m is the average magnitude of a weight.
- **A simple linear model can have adversarial examples if its input has sufficient dimensionality.**

The Boundary Tilting Perspective

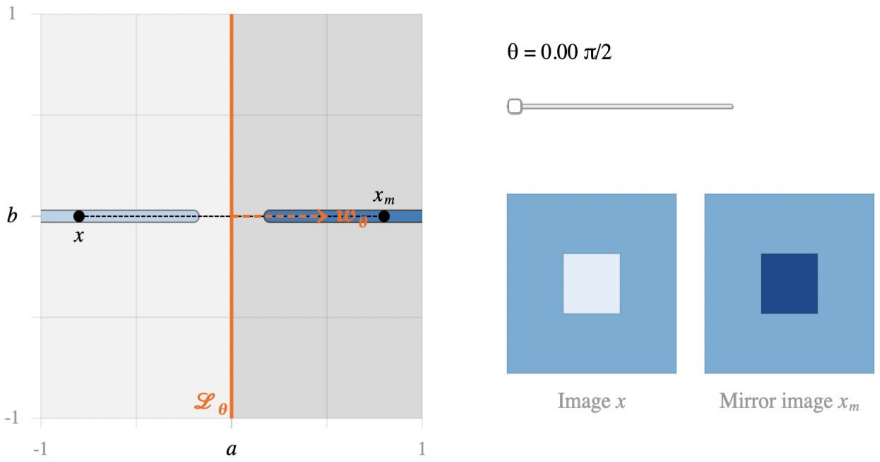


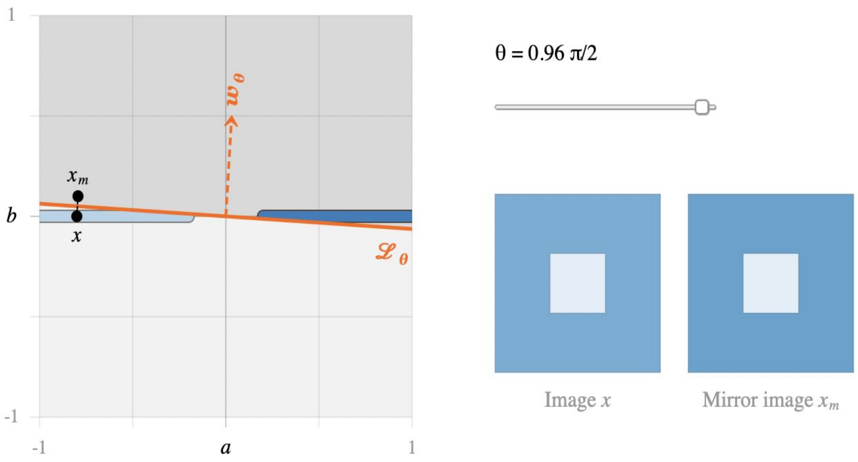
The Boundary Tilting Perspective



The line $\mathscr{L}_\theta$ defined by its normal weight vector $\boldsymbol{w}_\theta = (\cos\theta, \sin\theta)$ separates I and J for all θ in [0, π/2)

The Boundary Tilting Perspective



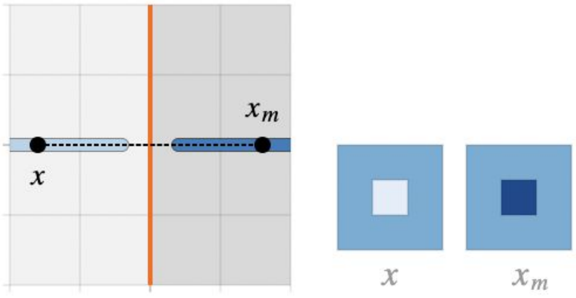


The Boundary Tilting Perspective

When $\theta = 0$:

$\mathcal{L}_\theta$ does not suffer from adversarial examples.

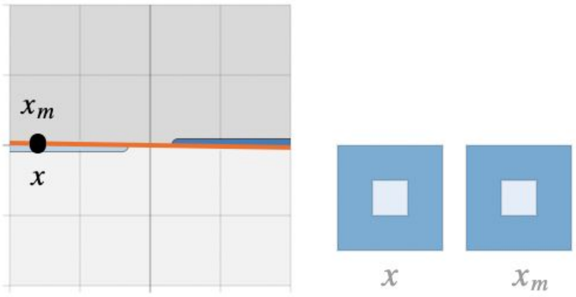
$\boldsymbol{x}$ is classified in I with high confidence and $\boldsymbol{x}_m$ is classified in J with high confidence, in agreement with human observers.



When $\theta \to \pi/2$:

$\mathcal{L}_\theta$ suffers from strong adversarial examples.

$\boldsymbol{x}$ is classified in I with high confidence and $\boldsymbol{x}_m$ is classified in J with high confidence, yet $\boldsymbol{x}_m$ is visually indistinguishable from $\boldsymbol{x}$.



The Boundary Tilting Perspective