Vulnerability of Deep Neural Networks

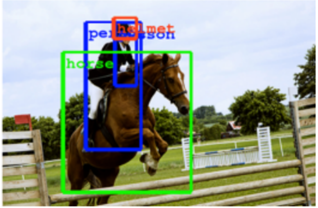
Shihua Zhang

November 17, 2021

Outline

Deep Neural Networks (DNNs)

DNNs are as good as humans at many tasks.



(Szegedy et al, 2014)

...recognizing objects and faces....

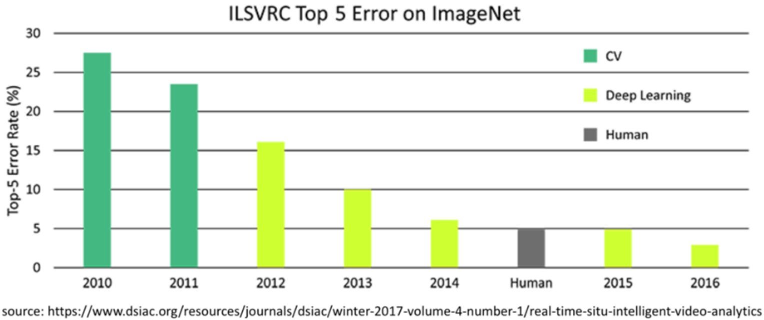
(Taigmen et al, 2013)

(Goodfellow et al, 2013)

...solving CAPTCHAS and reading addresses...

(Goodfellow et al, 2013)

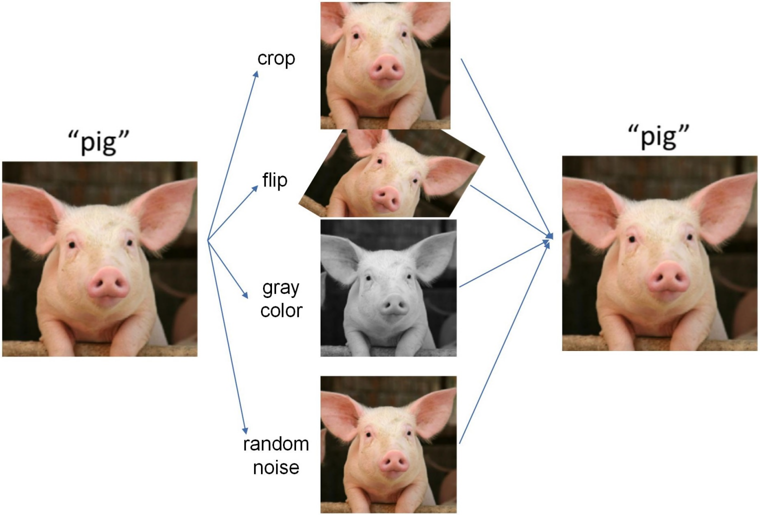
Beyond Human-Level Accuracy



Many architectures: VGG-Net, ResNet, DenseNet [5, 12].

DNNs Are Robust

DNNs are robust to common transformations: crop, flip, color distort, Gaussian noise and so on.



Deep neural networks are very powerful, but also can be vulnerable to adversarial attacks.

Outline

Adversarial Attack

Adversarial Attack: a method to generate adversarial examples.

Adversarial Attack

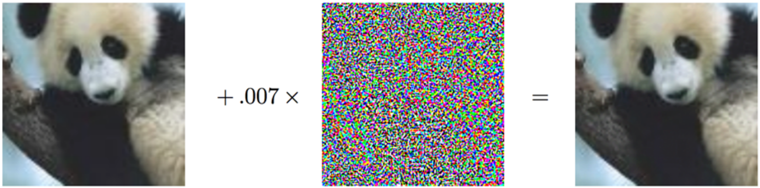
Adversarial Attack: a method to generate adversarial examples.

Adversarial Example: an instance with small, intentional feature perturbations that cause a learning model to make a false prediction.

Adversarial Attack

Adversarial Attack: a method to generate adversarial examples.

Adversarial Example: an instance with small, intentional feature perturbations that cause a learning model to make a false prediction.



$$\boldsymbol{x} + .007 \times \mathrm{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y)) = \boldsymbol{x} + \epsilon\mathrm{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$$

$\boldsymbol{x}$	$\mathrm{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$	$\boldsymbol{x} + \epsilon\mathrm{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$
"panda"	"nematode"	"gibbon"
57.7% confidence	8.2% confidence	99.3 % confidence

Threat to Safety of DNNs

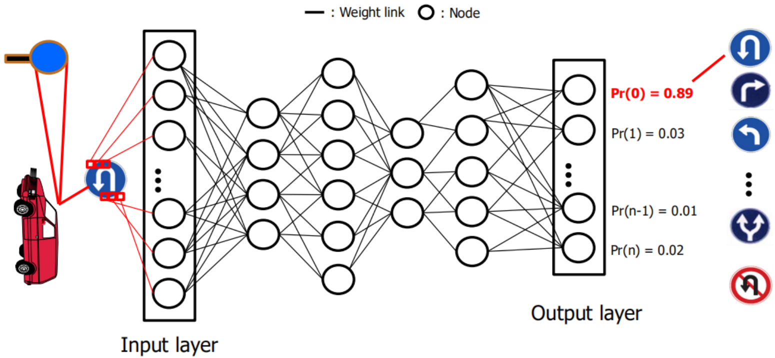
It is not difficult to fool a classifier

- The perturbation could be perceptually not noticeable
- Slightly modified data could lead to incorrect classification

Threat to Safety of DNNs

It is not difficult to fool a classifier

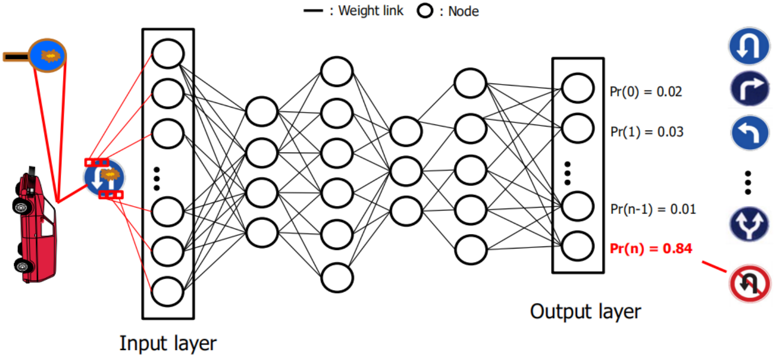
- The perturbation could be perceptually not noticeable
- Slightly modified data could lead to incorrect classification



Threat to Safety of DNNs

Adversarial Examples:

- Slightly modified data could lead to incorrect classification



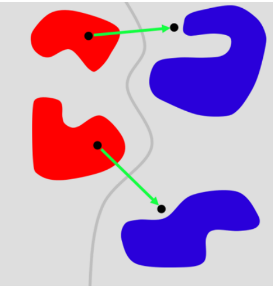
How to Generate Adversarial Examples?

Definition (Adversarial Attack)

Let $x_0 \in \mathbb{R}^d$ be a data point belong to class $\mathcal{C}_i$. Define a target class $\mathcal{C}_t$. An adversarial attack is a mapping $\mathcal{A} : \mathbb{R}^d \to \mathbb{R}^d$ such that the perturbed data

$$x = \mathcal{A}(x_0)$$

is misclassified as $\mathcal{C}_t$.

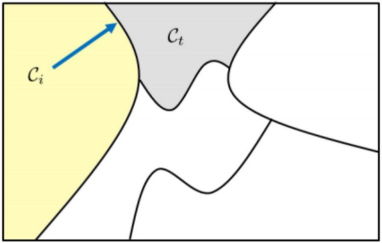


Targeted vs Untargeted Attack

The attack can be targeted or untargeted according to the choice [9, 13].

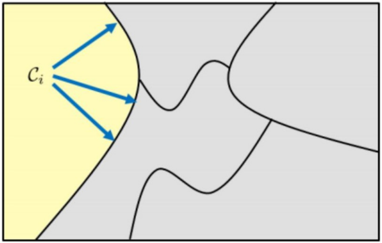
Targeted vs Untargeted Attack

The attack can be targeted or untargeted according to the choice [9, 13].



Targeted Attack:
[1]. The attack has to be specific from class i to class t.
[2]. The constraint set is

$$\Omega = \left\{ \boldsymbol{x} \mid \max_{j \neq t} \{g_j(\boldsymbol{x})\} - g_t(\boldsymbol{x}) \leq 0 \right\}$$



Untargeted Attack:
[1]. The attack vector can point to anywhere outside class i.
[2]. The constraint set is

$$\Omega = \left\{ \boldsymbol{x} \mid g_i(\boldsymbol{x}) - \min_{j \neq i} \{g_j(\boldsymbol{x})\} \leq 0 \right\}$$

White-box vs Black-box Attack

The attack can be white-box or block-box. It depends on your knowledge of the classifier [6, 7].

White-box vs Black-box Attack

The attack can be white-box or block-box. It depends on your knowledge of the classifier [6, 7].

White-box Attack	**Black-box Attack**
[1]. You know everything about the classifier. [2]. The constraint set is $\Omega = \left\{ \boldsymbol{x} \mid \max_{j \neq t} \{g_j(\boldsymbol{x})\} - g_t(\boldsymbol{x}) \leq 0 \right\}$ - where g(x) is the output of network w.r.t the input x.	[1]. You can only probe the classifier finite times. [2]. The constraint set is $\Omega = \left\{ \boldsymbol{x} \mid \max_{j \neq t} \{\widehat{g}_j(\boldsymbol{x})\} - \widehat{g}_t(\boldsymbol{x}) \leq 0 \right\}$ - where $\hat{g}$ is the best approximation you can get from the finite observations. - attacks can transfer among classifiers

Three Attack Forms

- We focus on targeted, white-box attack methods for simplicity.

Three Attack Forms

- We focus on targeted, white-box attack methods for simplicity.
- The three forms of attacks:
  - **Minimum Distance Attack**: Minimize the perturbation magnitude while accomplishing the attack objective
  - **Maximum Loss Attack**: Maximize the training loss while ensuring perturbation is controlled
  - **Regularization-based Attack**: Use regularization to control the amount of perturbation
- We will take linear classifier case as examples to gain insights.

Minimum Distance Attack

Definition (Minimum Distance Attack)

It finds a perturbed data x by solving the optimization

$$\begin{aligned} \underset{x}{\text{minimize}} \quad & \|x - x_0\| \\ \text{subject to} \quad & \max_{j \neq t} \{g_j(x)\} - g_t(x) \leqslant 0 \end{aligned}$$

where $\|\cdot\|$ can be any norm specified by the user.

Minimum Distance Attack

Definition (Minimum Distance Attack)

It finds a perturbed data x by solving the optimization

$$\begin{aligned} \underset{x}{\text{minimize}} \quad & \|x - x_0\| \\ \text{subject to} \quad & \max_{j \neq t} \{g_j(x)\} - g_t(x) \leqslant 0 \end{aligned}$$

where $\|\cdot\|$ can be any norm specified by the user.

- We aim to predict x as class $\mathcal{C}_t$.
- The constraint needs to be satisfied.
- It is desired to minimize the attack strength. This gives the objective.

Geometry: Attack as a Projection

Theorem (Minimum-Distance Attack as a Projection)

The minimum-distance attack via ℓ_2 is equivalent to the projection

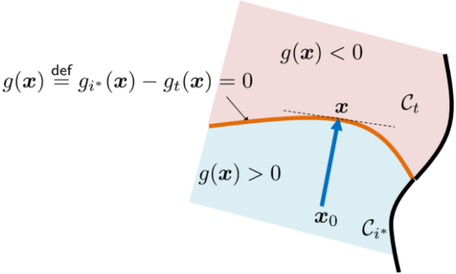
$$\begin{aligned} x^* &= \operatorname*{argmin}_{x \in \Omega} \|x - x_0\|^2, \quad \text{where } \Omega = \left\{ x \mid \max_{j \neq t} \{g_j(x)\} - g_t(x) \leqslant 0 \right\}, \\ &= \mathcal{P}_\Omega (x_0) \end{aligned} \tag{1}$$

Geometry: Attack as a Projection

Theorem (Minimum-Distance Attack as a Projection)

The minimum-distance attack via ℓ_2 is equivalent to the projection

$$\begin{aligned} x^* &= \underset{x \in \Omega}{\operatorname{argmin}} \|x - x_0\|^2, \quad \text{where } \Omega = \left\{ x \mid \max_{j \neq t} \{g_j(x)\} - g_t(x) \leqslant 0 \right\}, \\ &= \mathcal{P}_\Omega(x_0) \end{aligned} \tag{1}$$



Example: Binary Linear Classifier

In the binary linear case, the min-distance attack (ℓ_2-norm) becomes

1. Linear, we have

$$g_i(x) - g_t(x) = w^T x + w_0$$

2. Two classes: the constraint is simplified to

$$g_i(x) - g_t(x) \leqslant 0$$

- Thus, the attack becomes

$$\begin{aligned} \underset{x}{\text{minimize}} \quad & \|x - x_0\|^2 \\ \text{subject to} \quad & w^T x + w_0 = 0 \end{aligned}$$

Example: The ℓ_2

Theorem (Minimum ℓ_2-Norm Attack for Binary Linear Classifier)

The adversarial attack to a binary linear classifier is the solution of $\underset{x}{\text{minimize}}\ \|x - x_0\|^2$ subject to $w^T x + w_0 = 0$, which is given by

$$x^* = x_0 - \left(\frac{w^T x_0 + w_0}{\|w\|_2}\right)\frac{w}{\|w\|_2}.$$

- This is just finding the closest point to a hyperplane!
- $w/\|w\|_2$ is the normal direction $=$ best attack angle.
- $\frac{w^\top x_0 + w_0}{\|w\|_2}$ is the step size.

Example: The ℓ_2

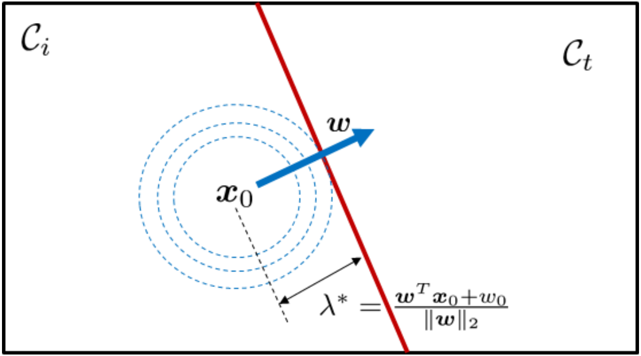


Figure: Geometry of minimum-distance attack for a two-class linear classifier with objective function $\|x - x_0\|^2$. The solution is a projection of the input x_0 onto the separating hyperplane of the classifier.

Example: The ℓ_∞ Solution

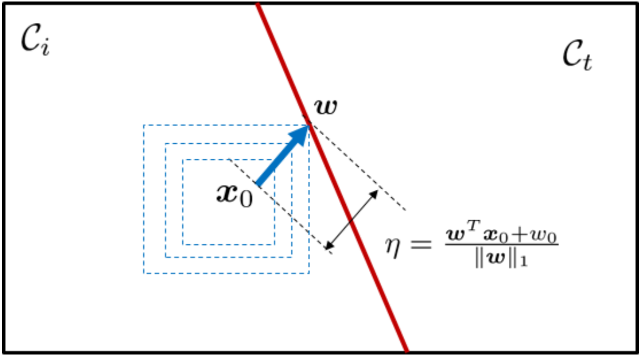
Theorem (Minimum Distance ℓ_∞-Norm Attack for Binary Linear Classifier)

The minimum distance ℓ_∞-norm attack for a binary linear classifier, i.e., $\underset{x}{\text{minimize}}\ \|x - x_0\|_\infty$ subject to $w^T x + w_0 = 0$ is given by

$$x = x_0 - \left(\frac{w^\top x_0 + w_0}{\|w\|_1}\right) \cdot \text{sign}(w).$$

- Search direction is $\text{sign}(w)$.
- This means ± 1 for every entry.
- In $2\mathrm{D}$, the search direction is $\pm 45^\circ$ or $\pm 135^\circ$.

Example: The ℓ_∞ Solution



- Is it the "optimal" direction? No.
- The fastest search direction is ℓ_2.
- η is larger to move x_0 to another class.

Maximum Loss Attack

Definition (Maximum Loss Attack)

It finds a perturbed data x by solving the optimization

$$\begin{array}{ll} \underset{x}{\text{maximize}} & g_t(x) - \max_{j \neq t} \{g_j(x)\} \\ \text{subject to} & \|x - x_0\| \leqslant \eta \end{array}$$

where $\|\cdot\|$ can be any norm specified by the user, and $\eta > 0$ denotes the attack strength.

- bound attack $\|x - x_0\| \leqslant \eta$
- make $g_t(x)$ as big as possible
- Thus, maximize $g_t(x) - \max_{j \neq t} \{g_j(x)\}$

Example: Binary Linear classification

- The problem is equivalent to

$$\begin{array}{ll} \underset{x}{\text{minimize}} & \max_{j \neq t} \{g_j(x)\} - g_t(x) \\ \text{subject to} & \|x - x_0\| \leqslant \eta \end{array}$$

- η is the maximum loss attack strength
- Want $g_t(x)$ to override $\max_{j \neq t} \{g_j(x)\}$

Example: Binary Linear classification

- The problem is equivalent to

$$\begin{array}{ll} \underset{x}{\text{minimize}} & \max_{j \neq t} \{g_j(x)\} - g_t(x) \\ \text{subject to} & \|x - x_0\| \leqslant \eta \end{array}$$

- η is the maximum loss attack strength
- Want $g_t(x)$ to override $\max_{j \neq t} \{g_j(x)\}$

- If you restrict to linear and only two classes, then

$$\underset{x}{\text{minimize}} \quad w^T x + w_0 \text{ subject to } \|x - x_0\| \leqslant \eta$$

- Solvable in closed-form.

Max-Loss Attack using ℓ_2-norm

- The problem is

$$\underset{r}{\text{minimize}} \quad w^T r + b_0 \text{ subject to } \|r\|_2 \leqslant \eta$$

- Cauchy inequality:

$$w^T r \geqslant -\|w\|_2 \|r\|_2 \geqslant -\eta \|w\|_2$$

- Claim: Lower bound of $w^T r$ is attained when $r = -\eta w / \|w\|_2$:

$$\begin{aligned} w^T r &= w^T \left(-\frac{\eta w}{\|w\|_2} \right) \\ &= -\eta \|w\|_2 \end{aligned}$$

- So the solution is $r = -\eta w / \|w\|_2$.

Regularization-based Attack

Definition (Regularization-based Attack)

It finds a perturbed data x by solving the optimization

$$\underset{x}{\text{minimize}} \quad \|x - x_0\| + \lambda \left(\max_{j \neq t} \{g_j(x)\} - g_t(x) \right)$$

where $\|\cdot\|$ can be any norm specified by the user, and $\lambda > 0$ is a regularization parameter.

- Combine the two parts via regularization
- By adjusting $(\epsilon, \eta, \lambda)$, all three will give the same optimal value.

Example: Binary Linear Classifier

Theorem (Regularization-based Attack for Binary Linear Classifier)

The regularization-based attack for a binary linear classifier generates the attack by solving

$$\underset{x}{\text{minimize}} \quad \frac{1}{2}\|x - x_0\|^2 + \lambda\left(w^T x + w_0\right)$$

of which the solution is given by

$$x = x_0 - \lambda w$$

- w is search direction.
- λ is the step size.

Summary

Three forms of adversarial attacks.

- Min-Distance Attack

$$\begin{aligned} \underset{x}{\text{minimize}} \quad & \|\boldsymbol{x} - \boldsymbol{x}_0\| \\ \text{subject to} \quad & \max_{j \neq t} \{g_j(\boldsymbol{x})\} - g_t(\boldsymbol{x}) \leqslant 0 \end{aligned}$$

- Max-Loss Attack

$$\begin{aligned} \underset{x}{\text{maximize}} \quad & g_t(\boldsymbol{x}) - \max_{j \neq t} \{g_j(\boldsymbol{x})\} \\ \text{subject to} \quad & \|\boldsymbol{x} - \boldsymbol{x}_0\| \leqslant \eta \end{aligned}$$

- Regularized Attack

$$\underset{\boldsymbol{x}}{\text{minimize}} \quad \|\boldsymbol{x} - \boldsymbol{x}_0\| + \lambda \left(\max_{j \neq t} \{g_j(\boldsymbol{x})\} - g_t(\boldsymbol{x}) \right)$$

Adversarial Defense

There are lots of strategies to defend the adversarial examples.

For example,

- Data Augmentation (e.g., dropout [1], mixup [14]).
- Feature Regularization [2, 8, 10].
- Adversarial Training [4].

Adversarial Defense

There are lots of strategies to defend the adversarial examples.

For example,

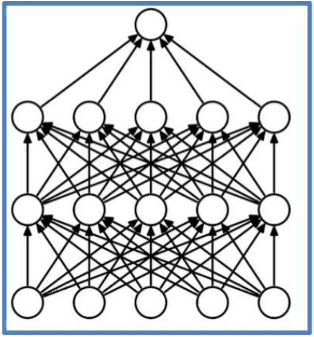
- Data Augmentation (e.g., dropout [1], mixup [14]).
- Feature Regularization [2, 8, 10].
- Adversarial Training [4].

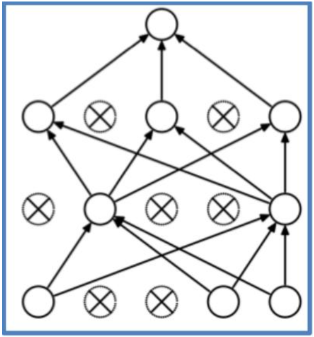
Adversarial training is an intuitive defense method against adversarial samples, which attempts to improve the robustness of a neural network by training it with adversarial samples.

Adversarial training is widely used and the most effective one.

Example: Dropout

- Randomly set some neurons and their connections to zeros.
- "Dropout" can be taken as methods of data augmentation and model ensemble





- With the dropout rate as its hyperparameter!

Example: Input Gradient Regularization

Consider the first-order Taylor expansion of the loss function,

$$\ell(x+\delta) - \ell(x) \approx g_\ell(x)^T \delta$$

where $g_\ell(x)$ denotes the gradients of loss function w.r.t. the input x.

- Input gradient regularization promotes smooth input gradients with fewer extreme values.
- Train neural networks by minimizing the input gradients[1]

$$\min \ell(W, x, y) + \lambda \|g_\ell(x)\|_2$$

[1] Ross, AS., Doshi-Velez, F. Improving the Adversarial Robustness and Interpretability of Deep Neural Networks by Regularizing their Input Gradients. AAAI (2018).

Example: Adversarial Training

Adversarial training [4] tried to solve a minimax problem,

$$\underset{\theta}{\text{minimize}} \quad \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\max_{\delta\in\mathcal{S}} L(\theta, x+\delta, y)\right]$$

- $\delta \in \mathcal{S}$ is the attack added to the input data x. y is the truth.
- $\mathcal{S}$ defines the set of allowable attacks, (e.g., ℓ_2 ball).
- The optimization can be done alternatively,
  - maximize of the loss $L(\theta, x+\delta, y)$ by searching for the most nasty attack δ.
  - minimize empirical risk over the training set $\mathcal{D}$.

Example: Adversarial Training

Adversarial training requires a significantly more complicated boundary and mitigate adversarial effects.

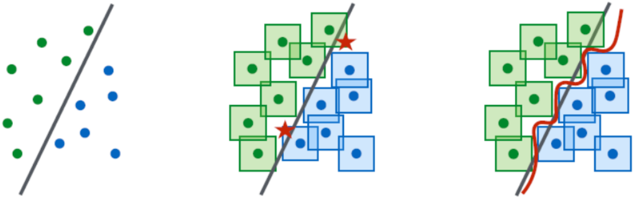


Figure: A conceptual illustration of standard vs. adversarial boundaries.

Outline

Are Adversarial Attacks Unavoidable?

Question: Is there any classifier that cannot be attacked?

Are Adversarial Attacks Unavoidable?

Question: Is there any classifier that cannot be attacked?

- No. All classifiers are adversarial vulnerable!

Isoperimetric Inequality

Definition 1 (ϵ-expansion)

The ϵ-expansion of a subset $A \subset \Omega$ w.r.t. distance metric d, denoted as $A(\epsilon, d)$, contains all points that are at most ϵ units away from A.

$$A(\epsilon, d) = \{x \in \Omega | d(x, y) \leqslant \epsilon \text{ for some } y \in A\}.$$

We simply write $A(\epsilon)$ when the distance metric is clear from context.

Lemma 1 (Isoperimetric inequality)

Consider a subset of the sphere $A \subset S^{n-1} \subset R^n$ with normalized measure $\mu_1(A) \geqslant 1/2$. When using the geodesic metric, the ϵ-expansion $A(\epsilon)$ is at least as large as the ϵ-expansion of a half sphere.

Isoperimetric Inequality

A special variant of the isoperimetric inequality first proved by Levy & Pellegrino (1951)

Lemma 2 (ϵ-expansion of half sphere)

The geodesic ϵ-expansion of a half sphere has normalized measure at least

$$1 - \left(\frac{\pi}{8}\right)^{\frac{1}{2}} \exp\left(-\frac{n-1}{2}\epsilon^2\right)$$

Existence of Adversarial Examples

Theorem 1 (Existence of Adversarial Examples)

Consider a classification problem with m classes, each distributed over the unit sphere S^{n-1} with density functions $\{\rho_c\}_{c=1}^m$. Choose a classifier function: $C : S^{n-1} \to \{1, 2, \cdots, m\}$ that partitions the sphere into disjoint measurable subsets. Define the following

- Let V_c denote the magnitude of the supremum of ρ_c relative to the uniform density. This can be written $V_c := s_{n-1} \cdot \sup_x \rho_c(x)$.
- Let $f_c = \mu_1\{x|C(x) = c\}$ be the fraction of the sphere labeled as c by classifier C. Choose some class c with $f_c \leqslant \frac{1}{2}$.

Sample a random data point x from ρ_c. Then with probability at least

$$1 - V_c \left(\frac{\pi}{8}\right)^{\frac{1}{2}} \exp\left(-\frac{n-1}{2}\epsilon^2\right) \tag{2}$$

x is misclassified by C, or x admits an ϵ-adversarial example in the geodesic distance.

Proof of Theorem 1

Proof. Choose a class c with $f_c \leqslant \frac{1}{2}$. Let $\mathcal{R} = \{x \mid \mathcal{C}(x) = c\}$ denote the region of the sphere labeled as class by $\mathcal{C}$, and let $\overline{\mathcal{R}}$ be its complement. $\overline{\mathcal{R}}(\epsilon)$ is the ϵ -expansion of $\overline{\mathcal{R}}$ in the geodesic metric. Because $\overline{\mathcal{R}}$ covers at least half the sphere, the isoperimetric inequality (Lemma 1) tells us that the ϵ-expansion is at least as great as the ϵ-expansion of a half sphere. We thus have

$$\mu_1[\bar{R}(\epsilon)] \geqslant 1 - \left(\frac{\pi}{8}\right)^{\frac{1}{2}} \exp\left(-\frac{n-1}{2}\epsilon^2\right)$$

Now, consider the set $\mathcal{S}_c$ of "safe" points from class c that are correctly classified and do not admit adversarial perturbations. A point is correctly classified only if it lies inside $\mathcal{R}$, and therefore outside of $\overline{\mathcal{R}}$. To be safe from adversarial perturbations, a point cannot lie within ϵ distance from the class boundary, and so it cannot lie within $\bar{R}(\epsilon)$. It is clear that the set $\mathcal{S}_c$ of safe points is exactly the complement of $\bar{R}(\epsilon)$.

Proof of Theorem 1

This set $\mathcal{S}_c$ has normalized measure

$$\mu_1\left[\mathcal{S}_c\right] \leqslant \left(\frac{\pi}{8}\right)^{\frac{1}{2}} \exp\left(-\frac{n-1}{2}\epsilon^2\right).$$

The probability of a random point lying in $\mathcal{S}_c$ is bounded above by the normalized supremum of ρ_c times the normalized measure $\mu_1\left[\mathcal{S}_c\right]$. This product is given by

$$V_c\left(\frac{\pi}{8}\right)^{\frac{1}{2}} \exp\left(-\frac{n-1}{2}\epsilon^2\right)$$

We then subtract this probability from 1 to obtain the probability of a point lying outside the safe region, and arrive at Eq. 3. □

Existence of Adversarial Examples

- The theorem tells that with probability at least

$$1 - V_c \left(\frac{\pi}{8}\right)^{\frac{1}{2}} \exp\left(-\frac{n-1}{2}\epsilon^2\right) \tag{3}$$

  the one of following will hold
  - The data x is originally misclassified, or
  - x can be attacked within an ϵ-ball.

- You can ignore the constant V_c.
- As the data dimension n grows, the probability will go to 1.
- So for large images, the probability of being attacked is high.
- A more general result without proof [3, 11].

Existence Result (Generally)

Theorem 2 (Adversarial examples on the cube)

Consider a classification problem with m classes, each distributed over the unit hypercube $[0,1]^n$ with density functions $\{\rho_c\}_{c=1}^m$. Choose a classifier function: $C : [0,1]^n \to \{1, 2, \cdots, m\}$ that partitions the hypercube into disjoint measurable subsets.

- Let U_c denote the supremum of ρ_c.
- Let f_c be the fraction of hypercube partitioned into class c by C.

Choose some class c with $f_c \leqslant \frac{1}{2}$, and select an ℓ_p-norm with $p > 0$. Define $p^* = \min(p, 2)$. Sample a random data point x from the class distribution ρ_c. Then with probability at least

$$1 - U_c \frac{\exp\left(-\pi n^{1-2/p^*} \epsilon^2\right)}{2\pi n^{1/2-1/p^*}} \tag{4}$$

one of the following conditions holds: x is misclassified by $\mathcal{C}$, or x has an adversarial example $\hat{x}$, with $\|x - \hat{x}\|_p \leqslant \epsilon$.

What do we learned?

Existence of Attack:

- The results above are only existence results.
- With high probability, there exists a direction which can almost certainly fool the classifier.
- This holds for all classifiers, as long as the dimension is high enough.
- Each perturbation pixel is small, but the sum can be big.

What do we learned?

Existence of Attack:

- The results above are only existence results.
- With high probability, there exists a direction which can almost certainly fool the classifier.
- This holds for all classifiers, as long as the dimension is high enough.
- Each perturbation pixel is small, but the sum can be big.

Can Random Noise Attack?

- Random noise cannot attack, especially for white-box.
- Probability of getting the correct attack direction is close to 0.
- Adversarial attacks are not common.

References I

A. Achille and S. Soatto.
Information dropout: Learning optimal representations through noisy computation.
IEEE Transactions on Pattern Analysis and Machine Intelligence, 40:2897–2905, 2018.

H. Drucker and Y. Le Cun.
Improving generalization performance using double backpropagation.
IEEE Transactions on Neural Networks, 3(6):991–997, 1992.

A. Fawzi, H. Fawzi, and O. Fawzi.
Adversarial vulnerability for any classifier.
In *NeurIPS*, 2018.

I. J. Goodfellow, J. Shlens, and C. Szegedy.
Explaining and harnessing adversarial examples.
CoRR, abs/1412.6572, 2015.

K. He, X. Zhang, S. Ren, and J. Sun.
Deep residual learning for image recognition.
2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 770–778, 2016.

J. Li, R. Ji, H. Liu, J. Liu, B. Zhong, C. Deng, and Q. Tian.
Projection & probability-driven black-box attack.
2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 359–368, 2020.

Y. Li, L. Li, L. Wang, T. Zhang, and B. Gong.
Nattack: Learning the distributions of adversarial examples for an improved black-box attack on deep neural networks.
In *ICML*, 2019.

References II

A. S. Rakin, Z. He, and D. Fan.
Parametric noise injection: Trainable randomness to improve deep neural network robustness against adversarial attack.
2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 588–597, 2019.

P. Rathore, A. Basak, S. H. Nistala, and V. Runkana.
Untargeted, targeted and universal adversarial attacks and defenses on time series.
2020 International Joint Conference on Neural Networks (IJCNN), pages 1–8, 2020.

A. Ross and F. Doshi-Velez.
Improving the adversarial robustness and interpretability of deep neural networks by regularizing their input gradients.
In *AAAI*, 2018.

A. Shafahi, W. R. Huang, C. Studer, S. Feizi, and T. Goldstein.
Are adversarial examples inevitable?
ArXiv, abs/1809.02104, 2019.

K. Simonyan and A. Zisserman.
Very deep convolutional networks for large-scale image recognition.
CoRR, abs/1409.1556, 2015.

A. Wu, Y. Han, Q. Zhang, and X. Kuang.
Untargeted adversarial attack via expanding the semantic gap.
2019 IEEE International Conference on Multimedia and Expo (ICME), pages 514–519, 2019.

C. Xie, J. Wang, Z. Zhang, Z. Ren, and A. Yuille.
Mitigating adversarial effects through randomization.
ArXiv, abs/1711.01991, 2018.