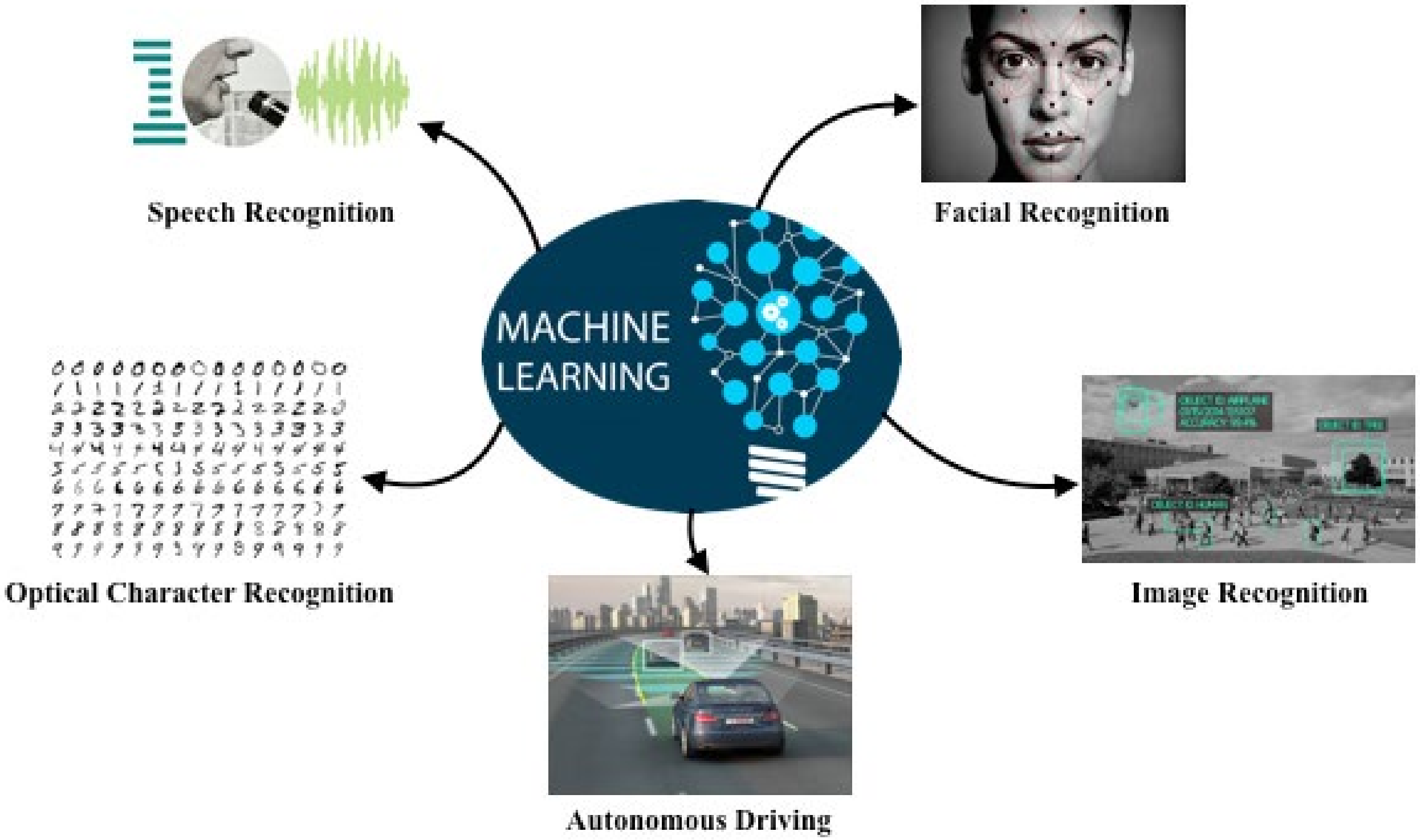
Adversarial Machine Learning

※ Attack algorithms

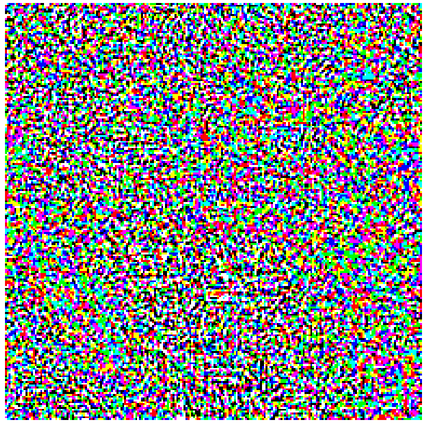
What is AML?


MACHINE LEARNING
Speech Recognition
Facial Recognition
Optical Character Recognition
Image Recognition
Autonomous Driving

Adversarial Examples

※ Attack algorithms

Classified as panda



Small adversarial noise

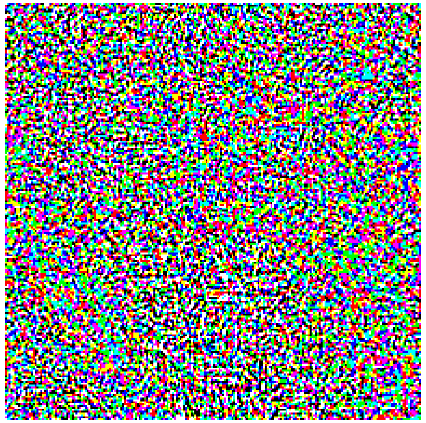
Classified as gibbon

?

Who cares panda?

Adversarial Examples

※ Attack algorithms



Small adversarial noise

?

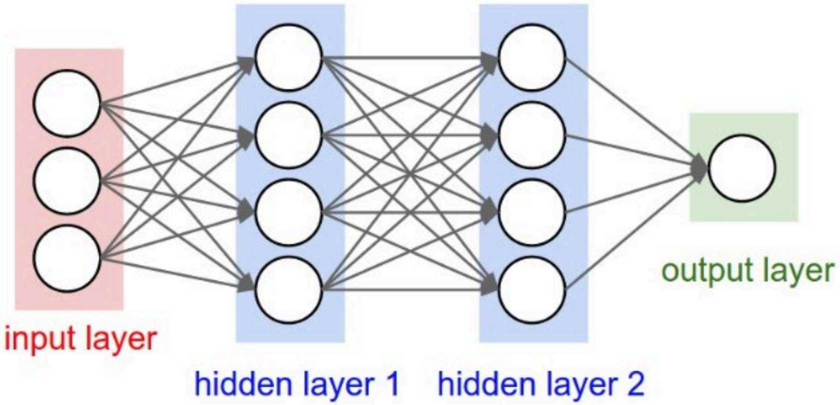
Outline

- Attack
  - Formulation
  - Distance metrics
- Attack algorithms
  - L-BFGS
  - Fast Gradient Sign
  - AdvGAN
  - One pixel attack

Attack

※ Attack algorithms

Original input



warplane

Attack: find a new input (similar to original input) but classified as another class t (untargeted or targeted)

Attacker knows the classifier

How to find adversarial examples

※ Attack algorithms

$$\begin{aligned} \textbf{minimize} \quad & \mathcal{D}(x, x+\delta) \\ \textbf{such that} \quad & C(x+\delta) = t \\ & x+\delta \in [0,1]^n \end{aligned}$$

distance between x and x+δ

x+δ is classified as target class t

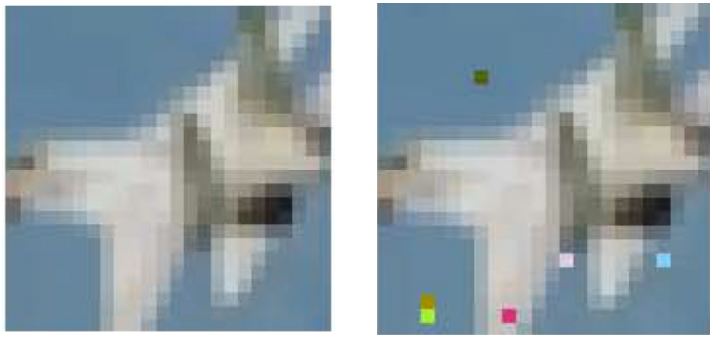
each element of x+δ in [0,1] (for a valid image)

Distance Metrics

※ Attack algorithms

Two images: x and x'

- L_0: measures the number of coordinates such that $x_i \neq x'_i$
  - corresponds to the number of pixels that have been changed in an image



- L_2: Euclidean distance

- L_∞: $\max(|x_1 - x'_1|, \dots, |x_n - x'_n|)$
  - measures maximum change to any of the elements

L-BFGS

※ Attack algorithms

$$\begin{aligned} \text{minimize} \quad & \|x - x'\|_2^2 \\ \text{such that} \quad & C(x') = l \\ & x' \in [0,1]^n \end{aligned}$$

$$\downarrow$$

$$\begin{aligned} \text{minimize} \quad & c \cdot \|x - x'\|_2^2 + \text{loss}_{F,l}(x') \\ \text{such that} \quad & x' \in [0,1]^n \end{aligned}$$

Initial formulation

$$\begin{aligned} \text{minimize} \quad & \mathcal{D}(x, x+\delta) \\ \text{such that} \quad & C(x+\delta) = t \\ & x + \delta \in [0,1]^n \end{aligned}$$

Note that these two are not equivalent optimization problems

SZEGEDY, C., ZAREMBA, W., SUTSKEVER, I., BRUNA, J., ERHAN, D., GOODFELLOW, I., AND FERGUS, R. Intriguing properties of neural networks. *ICLR* (2013

Fast Gradient Sign

※ Attack algorithms

$$x' = x - \epsilon \cdot \mathbf{sign}(\nabla \mathrm{loss}_{F,t}(x))$$

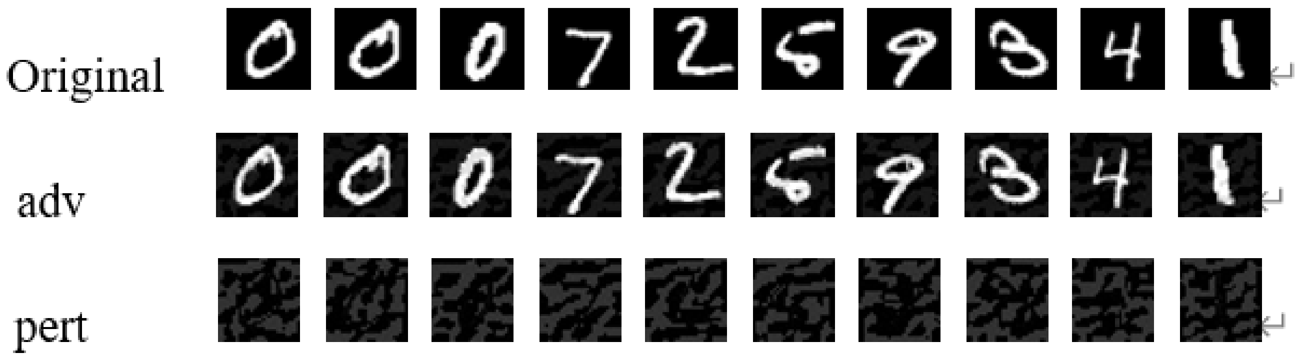
ϵ is chosen to be sufficiently small so as to be undetectable

fast rather than optimal

GOODFELLOW, I. J., SHLENS, J., AND SZEGEDY, C. Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572 (2014).

Fast Gradient Sign

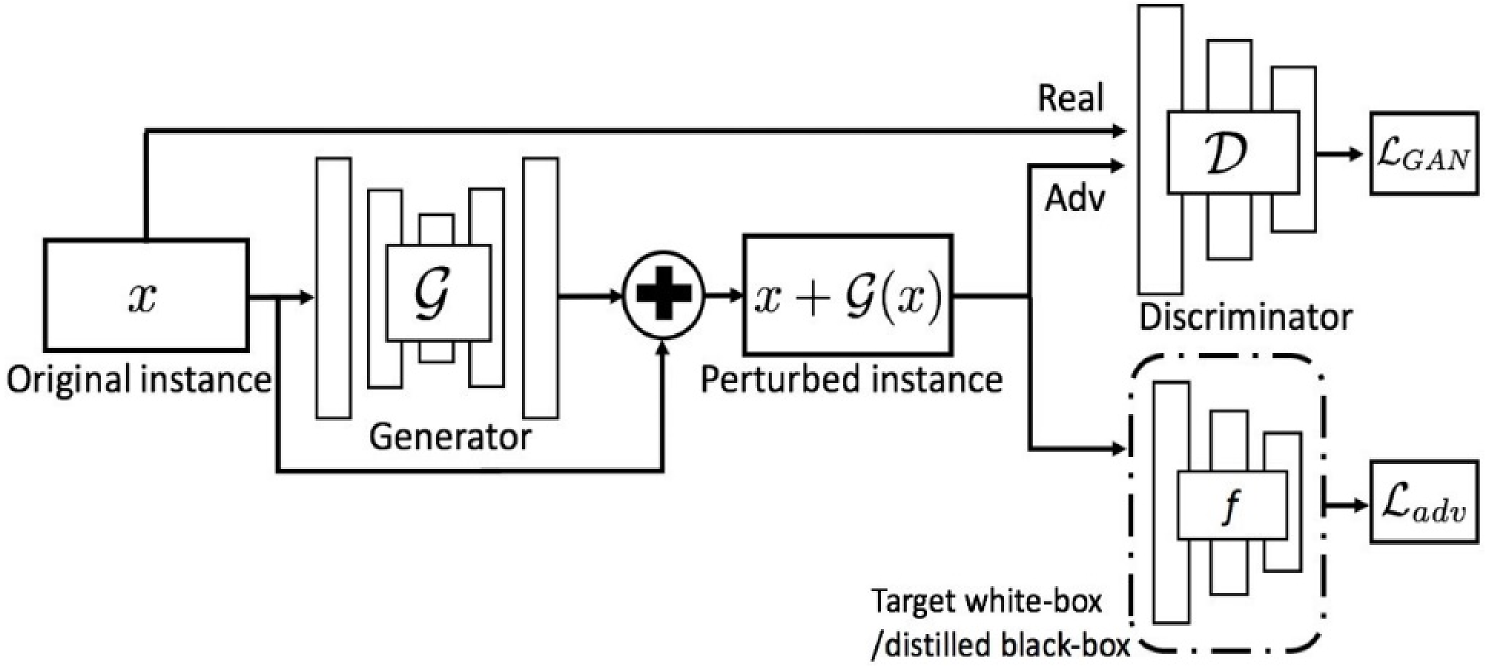
※ Attack algorithms



GOODFELLOW, I. J., SHLENS, J., AND SZEGEDY, C. Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572 (2014).

AdvGAN

※ Attack algorithms



$$\mathcal{L}_{\text{GAN}} = \mathbb{E}_x \log \mathcal{D}(x) + \mathbb{E}_x \log(1 - \mathcal{D}(x + \mathcal{G}(x))).$$

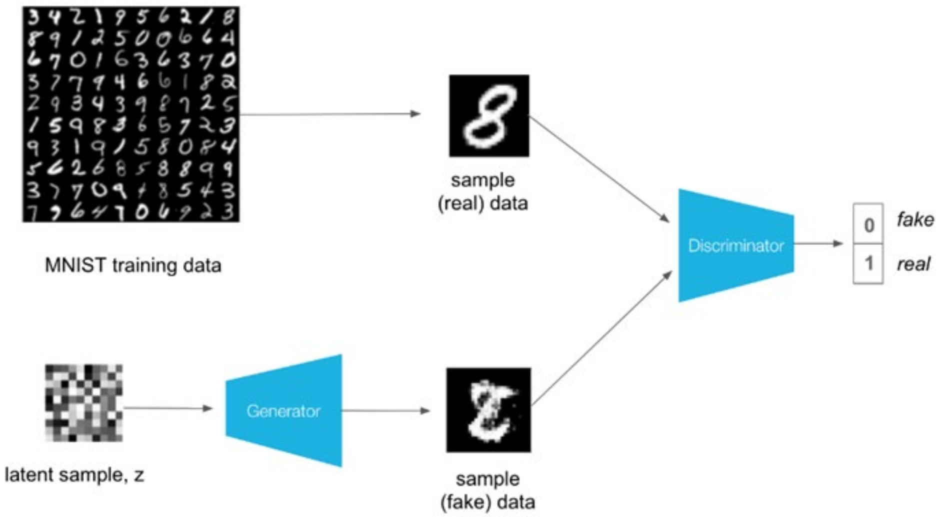
$$\mathcal{L}^f_{\text{adv}} = \mathbb{E}_x \ell_f(x + \mathcal{G}(x), t),$$

$$\mathcal{L}_{\text{hinge}} = \mathbb{E}_x \max(0, \|\mathcal{G}(x)\|_2 - c),$$

$$\mathcal{L} = \mathcal{L}^f_{\text{adv}} + \alpha \mathcal{L}_{\text{GAN}} + \beta \mathcal{L}_{\text{hinge}},$$

AdvGAN

※ Attack algorithms



Untargeted

Targeted

One pixel attack

※ Attack algorithms

$$\begin{aligned} \underset{e(\mathbf{x})^*}{\text{maximize}} \quad & f_{adv}(\mathbf{x} + e(\mathbf{x})) \\ \text{subject to} \quad & \|e(\mathbf{x})\| \leq L \end{aligned}$$

$\longrightarrow$

$$\begin{aligned} \underset{e(\mathbf{x})^*}{\text{maximize}} \quad & f_{adv}(\mathbf{x} + e(\mathbf{x})) \\ \text{subject to} \quad & \|e(\mathbf{x})\|_0 \leq d, \end{aligned}$$

modify a part of all dimensions

modify d dimensions

One pixel attack

Method	Success rate	Confidence	Number of pixels	Network
Our method	35.20%	60.08%	1 (0.098%)	NiN
Our method	31.40%	53.58%	1 (0.098%)	VGG
LSA[15]	97.89%	72%	33 (3.24%)	NiN
LSA[15]	97.98%	77%	30 (2.99%)	VGG
FGSM[11]	93.67%	93%	1024 (100%)	NiN
FGSM[11]	90.93%	90%	1024 (100%)	VGG

TABLE IX
COMPASSION OF NON-TARGETED ATTACK EFFECTIVENESS BETWEEN THE PROPOSED METHOD AND TWO PREVIOUS WORKS. THIS SUGGESTS THAT ONE PIXEL IS ENOUGH TO CREATE ADVERSARIAL IMAGES FROM MOST OF THE NATURAL IMAGES.