EXPLAINING AND HARNESSING ADVERSARIAL EXAMPLES

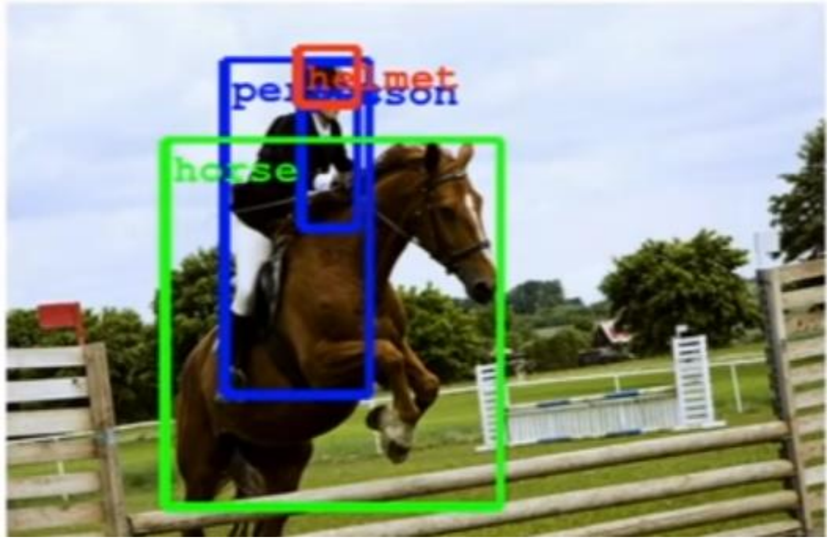
Ian J. Goodfellow, Jonathon Shlens, Christian Szegedy

ICLR’ 15

Dohyun Kim

Introduction

❖ Since 2013, deep neural networks have matched human performance



(Szegedy et. al., 2014)

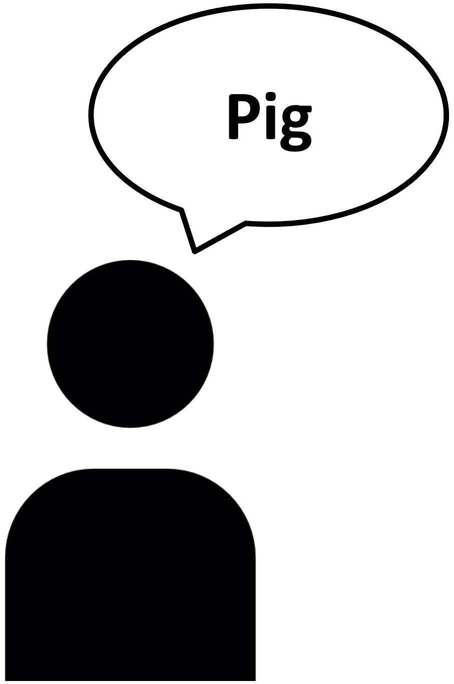
(Taigman et. al., 2013)

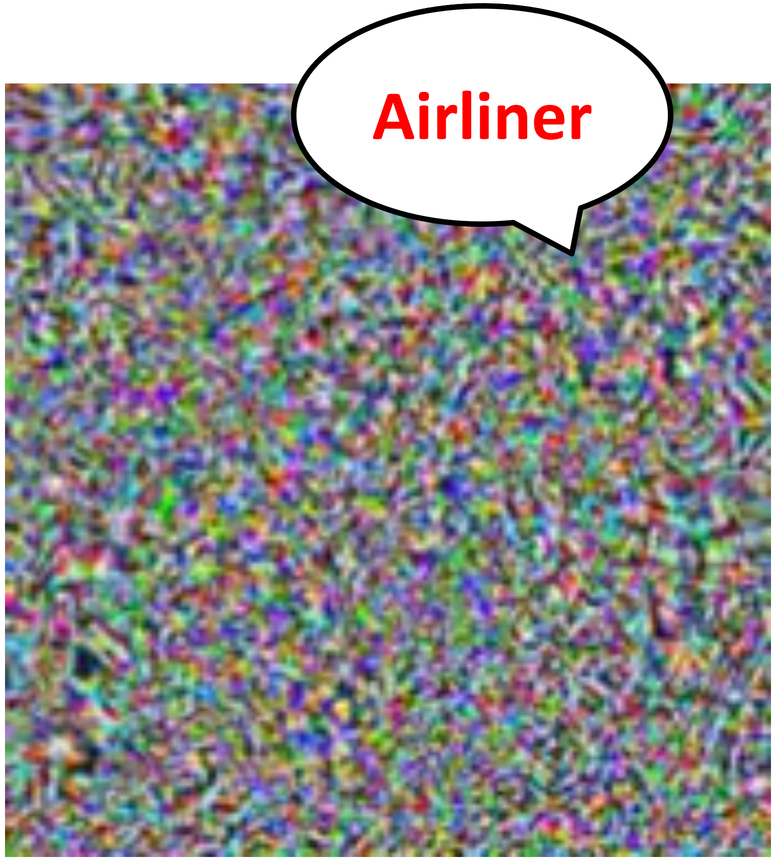
recognizing objects and faces

(Goodfellow et. al., 2013)

solving CAPTCHA

Adversarial Example





Adversarial Example

$\boldsymbol{x}$

"panda"
57.7% confidence

$+ .007 \times$

$\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$

"nematode"
8.2% confidence

$=$

$\boldsymbol{x} + \epsilon\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$

"gibbon"
99.3 % confidence

$=$

Related Works

- “Several machine learning models, including state-of-the-art neural networks, are vulnerable to **adversarial examples**”

Intriguing properties of neural networks

Christian Szegedy
Google Inc.

Wojciech Zaremba
New York University

Ilya Sutskever
Google Inc.

Joan Bruna
New York University

Dumitru Erhan
Google Inc.

Ian Goodfellow
University of Montreal

Rob Fergus
New York University
Facebook Inc.

Contribution

- The authors provided the first analysis of the causes of the adversarial examples
  - linear behavior in high-dimensional spaces

- Design a fast method of generating adversarial examples using linearity
  - Fast gradient sign method (FGSM)

- Adversarial training can provide an additional regularization benefit beyond that provided by using dropout alone

Linear Explanation of Adversarial Examples

Adversarial input

Small enough to be ignored

$$\tilde{\boldsymbol{x}} = \boldsymbol{x} + \boldsymbol{\eta} \qquad ||\boldsymbol{\eta}||_\infty < \epsilon.$$

❖ Consider the dot product between a weight vector and adversarial example

$$\boldsymbol{w}^\top \tilde{\boldsymbol{x}} = \boldsymbol{w}^\top \boldsymbol{x} + \boldsymbol{w}^\top \boldsymbol{\eta}.$$

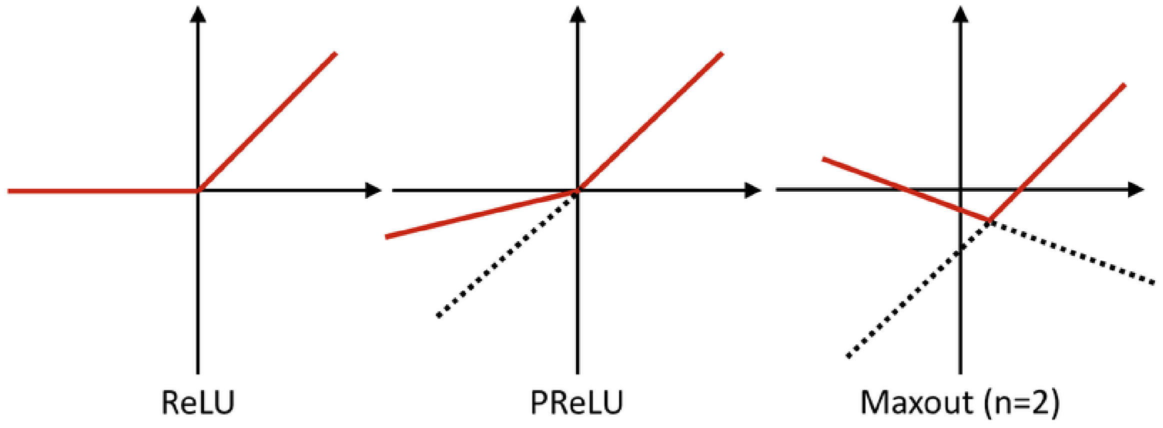
❖ Can maximize the increase when $\eta = \text{sign}(\boldsymbol{w})$

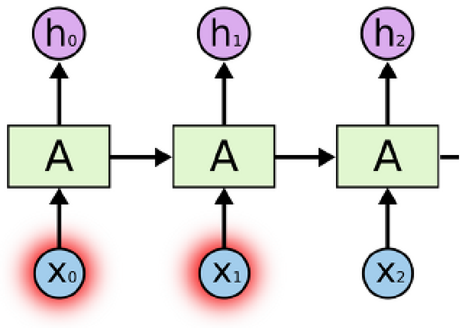
❖ Let $\eta = \epsilon \text{sign}(\omega)$, $\sum_i^n |\omega_i| = m$ for $\omega = [\omega_1\ \omega_2\ \cdots\ \omega_n]$

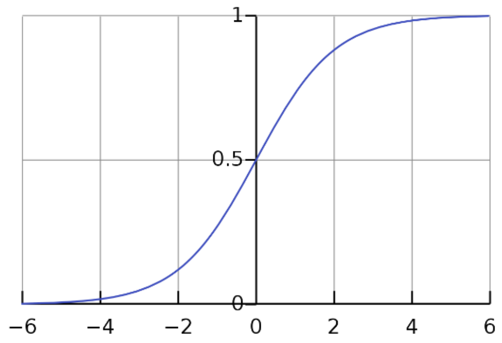
❖ $$\omega^{\mathrm{T}}\eta = [\omega_1\ \omega_2\ \cdots\ \omega_n]\begin{bmatrix}\eta_1\\ \eta_2\\ \vdots\\ \eta_n\end{bmatrix} = \omega_1\eta_1 + \omega_2\eta_2 + \cdots + \omega_n\eta_n = \epsilon m n$$

Linear Perturbation of Nonlinear Models

❖ Existing models such as LSTMs, ReLUs, maxout networks are designed to behave in linear ways



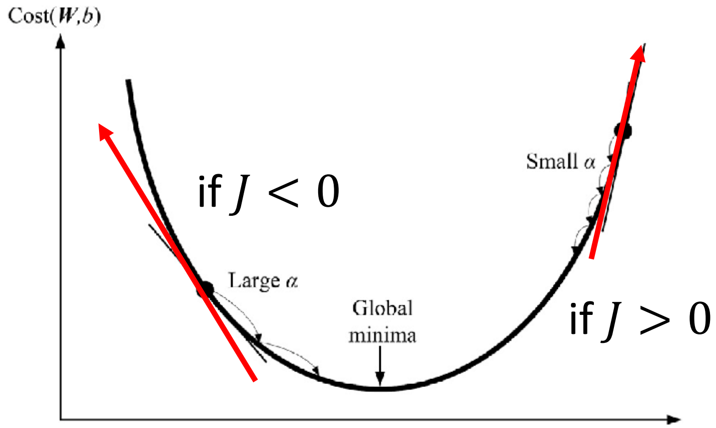


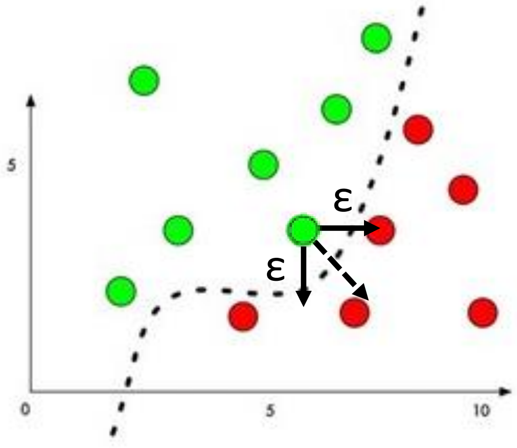


Linear Perturbation of Nonlinear Models

- θ: parameters of a model
- x: input to the model
- y: targets associated with x
- $J(\theta, x, y)$: cost used to train the neural network

$$\tilde{\boldsymbol{x}} = \boldsymbol{x} + \boldsymbol{\eta} \qquad \boldsymbol{\eta} = \epsilon \operatorname{sign}\left(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y)\right).$$





Adversarial Example Results

- FGSM causes a wide variety of models to misclassify their inputs

Model / Dataset	Error Rate (Adversarial)	Confidence
Shallow softmax Classifier / MNIST	99.9%	79.3%
Maxout network / MNIST	89.4%	97.6%
Conv maxout network / CIFAR-10	87.15%	96.6%

Adversarial Training of Deep Networks

- Training on a mixture of adversarial and clean examples, a neural network could be regularized

- θ: parameters of a model
- x: input to the model
- y: targets associated with x
- $J(\theta, x, y)$: cost used to train the neural network

$$\tilde{J}(\boldsymbol{\theta}, \boldsymbol{x}, y) = \alpha J(\boldsymbol{\theta}, \boldsymbol{x}, y) + (1 - \alpha) J(\boldsymbol{\theta}, \boldsymbol{x} + \epsilon \text{sign}\left(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y)\right)\right).$$

Adversarial Training Results

- Average confidence on a misclassified example was 81.4%

	w/o Adversarial	w/ Adversarial
Error rate	0.94%	0.84%
Adversarial error rate	89.4%	17.9%

Summary

- Adversarial examples can be explained as a property of high-dimensional dot products
- Introduced a fast method for generating adversarial examples
- Demonstrated that adversarial training can result in regularization; even further regularization than dropout
- Models that are easy to optimize are easy to perturb

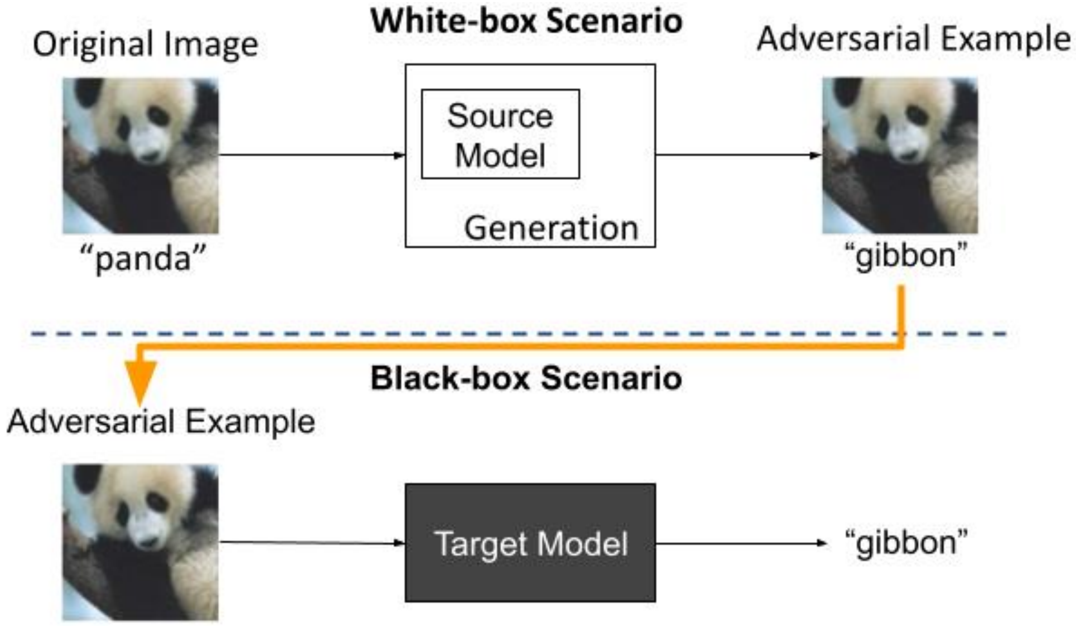
Future Works

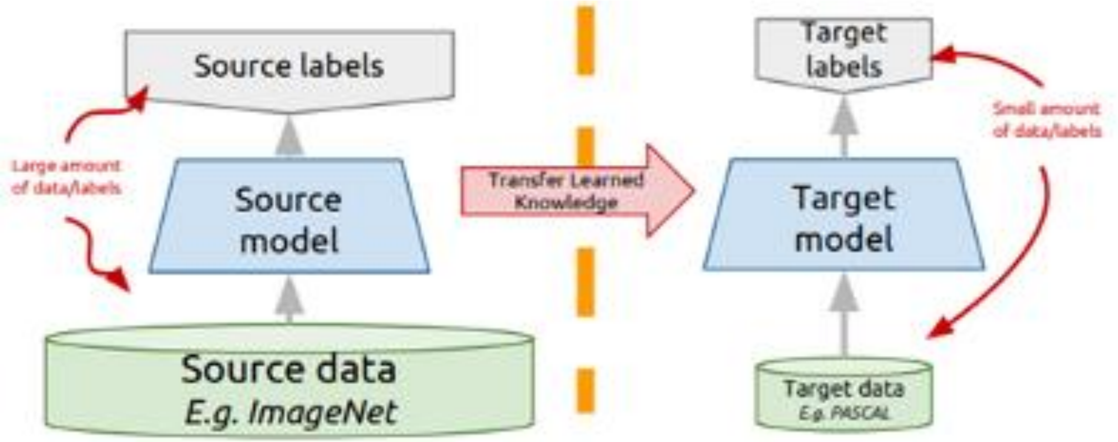
- Adversarial attacks have been developed in several ways
  - Image recognition adversarial attack
    - DeepFool, JSMA, FGSM, FGVM, IGSM, etc.
  - Audio adversarial attack
    - Audio Adversarial Examples: Targeted Attacks on Speech-to-Text
  - Physical-world adversarial attack
    - Robust Physical-World Attacks on Deep Learning Visual Classification

Q&A

- This attack is a white box attack. For grey or black box attack, what should the attacker consider more? (고우영)

- A) Adversarial attack has **transferable property**





Q&A

- Q) What is the most concrete defense method against adversarial attack? (한상구)

- A) There is no concrete defense method.
  - Adding noise at training time, error correcting codes, removing perturbation with an autoencoder, generative pretraining, confidence-reducing perturbation at test time, defensive distillation, etc.

- Machine learning techniques have built a **Potemkin village**
  - Shallow decision boundaries instead of actual underlying truths

Q&A

- Q) I believe a model is vulnerable if there's any adversarial input, as you can forge a lot of things. How much are those adversarial cases effective in real world? (정현식)

- Application of image classification
  - fingerprint authentication, face recognition, social media platform, healthcare industry, camera in automobile, etc.

- Adversarial attack in image classification is not so **attrac**
  - Is adversarial attack better than other attacks?