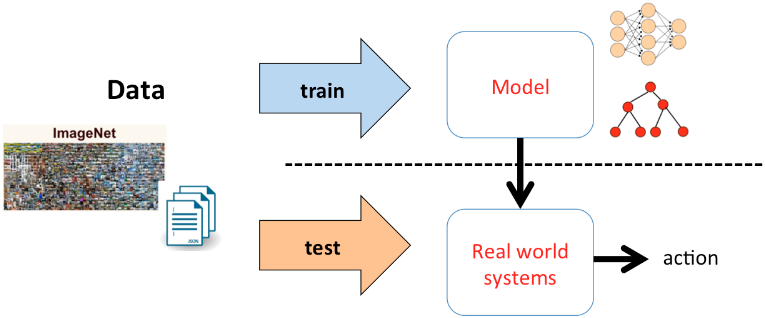
STOR566: Introduction to Deep Learning

Lecture 14: Adversarial Attack

Yao Li
UNC Chapel Hill

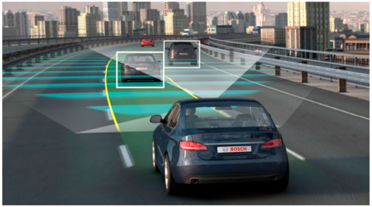
Oct 11, 2022

Machine Learning Systems



Testing: Robustness and Safety

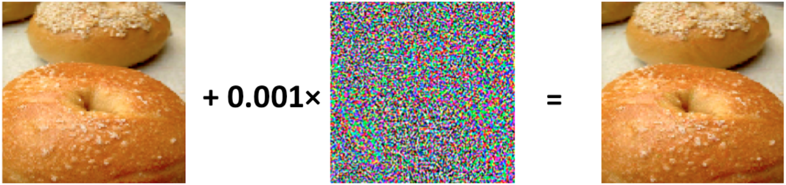
ML systems need to interact with real world



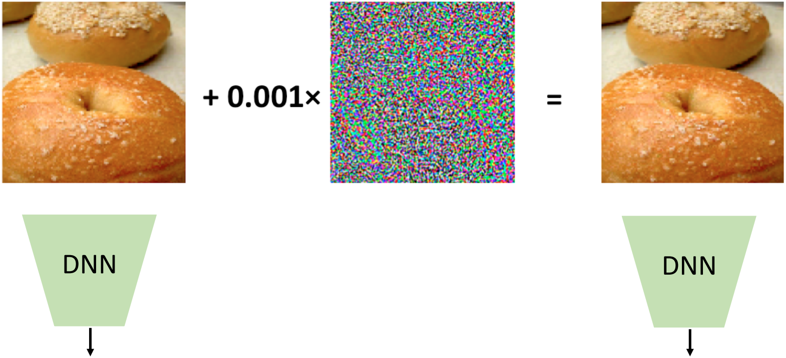
- Robustness and Safety

Adversarial Examples

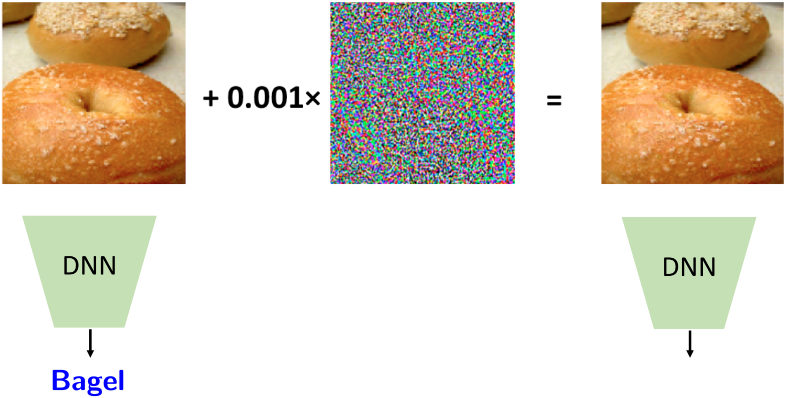
Adversarial Examples



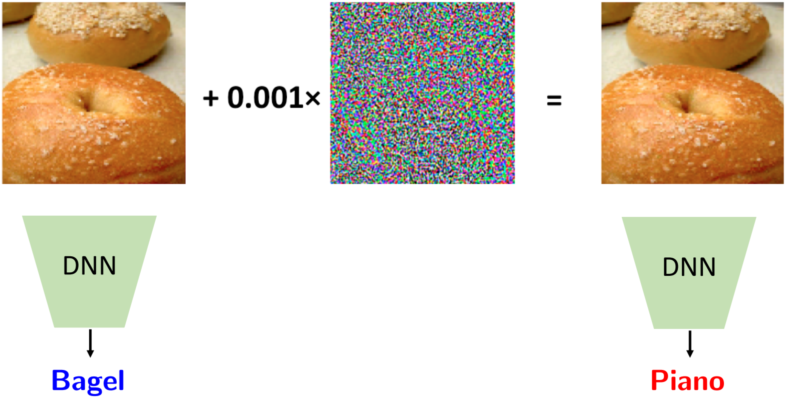
Adversarial Examples



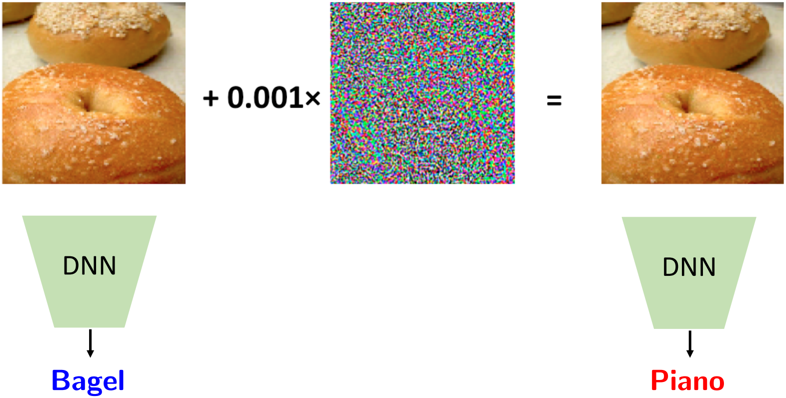
Adversarial Examples



Adversarial Examples



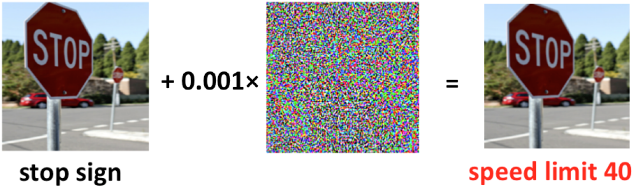
Adversarial Examples



A carefully crafted adversarial example can easily fool a deep network

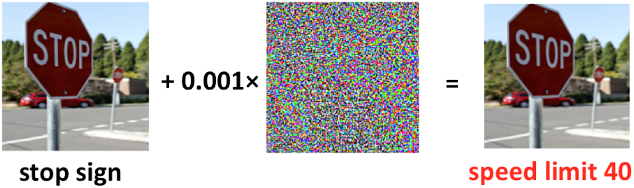
Adversarial Examples

- Robustness is critical in real systems



Adversarial Examples

- Robustness is critical in real systems



Not safe!

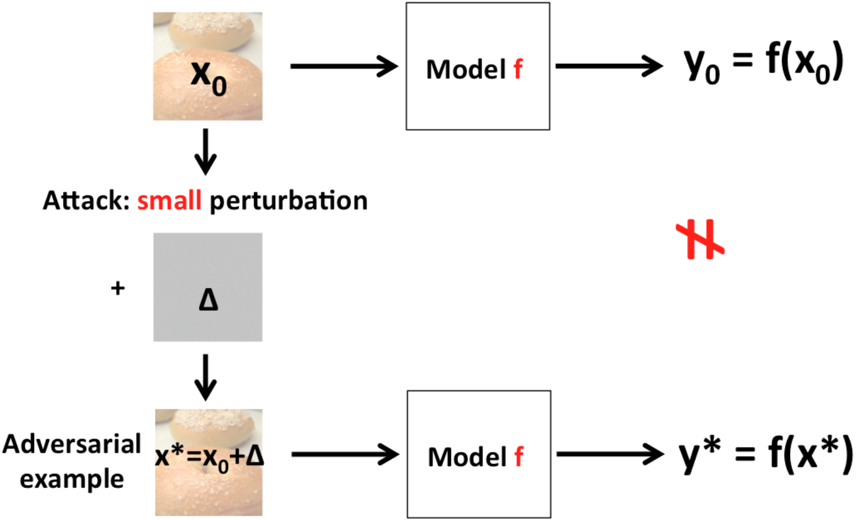
Adversarial Example in NLP

BAE attack on IMDB	
Groundtruth Label changed : Positive $\rightarrow$ Negative	
Original	This film offers many delights and surprises.
Attacked	This beautiful movie offers many pleasant delights and surprises.

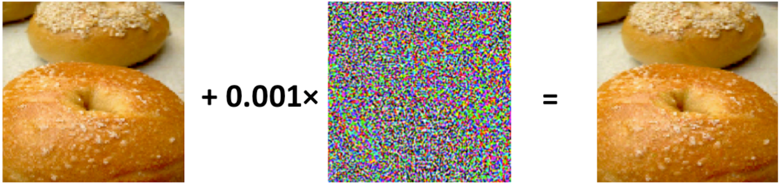
Table: BERT-based Adversarial Examples for Text Classification on IMDB dataset. (Replacements: Red, Inserts: Blue)

Adversarial Attack

Notations and Attack Procedure



Type of Attacks



Bagel y^*

1. Untargeted attack: $y^* \neq$ Bagel
2. Targeted attack: For target class $t =$ Piano, the attacker wants $y^* =$ Piano

Gradient-based Attack

Attack as Optimization

Attack as an optimization problem:

- Given prediction model with fixed parameter $\boldsymbol{\theta}$
- $(\boldsymbol{x}_0, y_0)$: input image and label

$$\boldsymbol{\delta} = \arg\max_{\boldsymbol{\delta} \in \mathcal{S}} L(\boldsymbol{\theta}, \boldsymbol{x}_0 + \boldsymbol{\delta}, y_0)$$
$$\boldsymbol{x}^* = \boldsymbol{x}_0 + \boldsymbol{\delta}$$

L: loss function training the classifier

$\boldsymbol{\delta}$: adversarial perturbation

$\mathcal{S} \in \mathbb{R}^d$: allowed perturbation set, usually chosen to be $\{\boldsymbol{\delta} | \|\boldsymbol{\delta}\|_\infty \leq \epsilon\}$

FGSM and PGD

Fast gradient sign method (FGSM):

- One-step gradient ascent:

$$\boldsymbol{x}^* = \boldsymbol{x}_0 + \epsilon \cdot \mathrm{sign}\Big(\nabla_{\boldsymbol{x}} L(\boldsymbol{\theta}, \boldsymbol{x}_0, y)\Big) \tag{1}$$

Projected Gradient Descent Attack (PGD):

- Multiple-steps gradient ascent:

$$\boldsymbol{x}^{t+1} = \Pi_\epsilon \left\{ \boldsymbol{x}^t + \alpha \cdot \mathrm{sign}\Big(\nabla_{\boldsymbol{x}} L(\boldsymbol{\theta}, \boldsymbol{x}^t, y)\Big), \boldsymbol{x}_0 \right\} \tag{2}$$

$\Pi_\epsilon(\cdot, \boldsymbol{x}_0)$: projection to the set $\{\boldsymbol{x} | \|\boldsymbol{x} - \boldsymbol{x}_0\|_\infty \leq \epsilon\}$

α: step size

Carlini and Wagner Attack (C&W Attack)

Given model $f(\cdot)$ and input image and label $(\boldsymbol{x}_0, y_0)$
Craft adversarial example by solving

$$\arg\min_{\boldsymbol{x}} \ \|\boldsymbol{x} - \boldsymbol{x}_0\|^2 + \lambda \cdot g(\boldsymbol{x})$$

(Carlini et al., 2017)

C&W Attack

Given model $f(\cdot)$ and input image and label $(\boldsymbol{x}_0, y_0)$
Craft adversarial example by solving

$$\arg\min_{\boldsymbol{x}} \; \|\boldsymbol{x} - \boldsymbol{x}_0\|^2 + \lambda \cdot g(\boldsymbol{x})$$

- $\|\boldsymbol{x} - \boldsymbol{x}_0\|^2$: the distortion

(Carlini et al., 2017)

C&W Attack

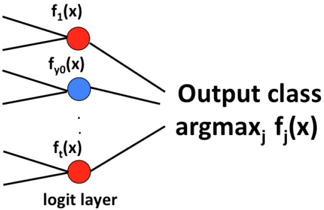
Given model $f(\cdot)$ and input image and label $(\boldsymbol{x}_0, y_0)$
Craft adversarial example by solving

$$\arg\min_{\boldsymbol{x}} \; \|\boldsymbol{x} - \boldsymbol{x}_0\|^2 + \lambda \cdot g(\boldsymbol{x})$$

- $\|\boldsymbol{x} - \boldsymbol{x}_0\|_2^2$: the distortion
- $g(\boldsymbol{x})$: loss to measure the successfulness of attack
- $\lambda \geq 0$: controls the trade-off

(Carlini et al., 2017)

Untargeted attack

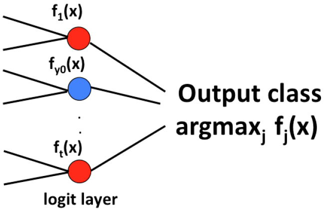


- $g(\boldsymbol{x})$: loss to measure the successfulness of attack

Untargeted attack: success if $\arg\max_j f_j(\boldsymbol{x}) \neq y_0$

$$g(\boldsymbol{x}) = \max\{f_{y_0}(\boldsymbol{x}) - \max_{j \neq y_0} f_j(\boldsymbol{x}), 0\}$$

Targeted attack



- $g(\boldsymbol{x})$: loss to measure the successfulness of attack

  Targeted attack: success if $\arg\max_j f_j(\boldsymbol{x}) = t$

$$g(\boldsymbol{x}) = \max\{\max_{j \neq t} f_j(\boldsymbol{x}) - f_t(\boldsymbol{x}), 0\}$$

Gradient-Based (White-box) Setting

- Model (network structure and weights) is revealed to attacker
  $\Rightarrow$ gradient of $g(\boldsymbol{x})$ (or $\nabla_{\boldsymbol{x}} L(\boldsymbol{\theta}, \boldsymbol{x}^t, y)$) can be computed by back-propagation
  $\Rightarrow$ attacker searches for $\boldsymbol{x}^*$ by gradient descent
- Distortion can be measured by other norms:
  (e.g., ℓ_∞, Elastic net, ...)
- Black-box setting: Only part of the information is available

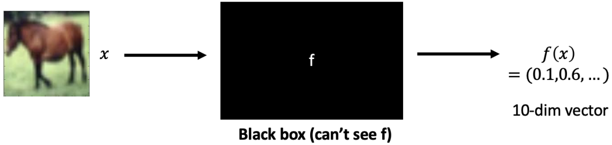
Score-based Attack

Score-based Setting

- In practice, the deep network parameters are not revealed to attackers

  cannot compute gradient

- Score-based setting: Attacker can query the model and get the score vector



- Example: CIFAR10 image classification

  $f(\boldsymbol{x}) \in \mathbb{R}^{10}$, $f(\boldsymbol{x})_i$: predicted score of class i

Zeroth Order Optimization based Attack (ZOO)

- We can solve the following problem without access to parameters:

$$\arg\min_{\boldsymbol{x}} \ \|\boldsymbol{x} - \boldsymbol{x}_0\|^2 + \lambda \cdot g(\boldsymbol{x})$$

- Estimate gradient:

$$\hat{g}_i \approx \frac{\partial f(\boldsymbol{x})}{\partial x_i} \approx \frac{f(\boldsymbol{x} + \delta \boldsymbol{e}_i) - f(\boldsymbol{x} - \delta \boldsymbol{e}_i)}{2\delta}$$

Zeroth Order Optimization based Attack (ZOO)

- We can solve the following problem without access to parameters:

$$\arg\min_{\boldsymbol{x}} \ \|\boldsymbol{x} - \boldsymbol{x}_0\|^2 + \lambda \cdot g(\boldsymbol{x})$$

- Estimate gradient:

$$\hat{g}_i \approx \frac{\partial f(\boldsymbol{x})}{\partial x_i} \approx \frac{f(\boldsymbol{x} + \delta \boldsymbol{e}_i) - f(\boldsymbol{x} - \delta \boldsymbol{e}_i)}{2\delta}$$

- zero-order gradient descent

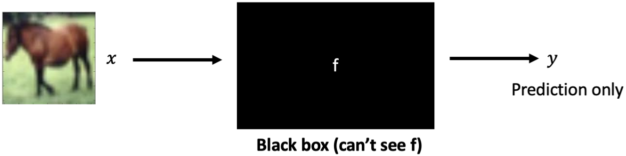
$$\boldsymbol{x} \leftarrow \boldsymbol{x} - \eta \begin{pmatrix} \hat{g}_1 \\ \vdots \\ \hat{g}_d \end{pmatrix}$$

(Chen et al., 2017)

Decision-based Attack

Decision-based Setting

- In practice, the attacker may only have access to the predicted label
- Decision-based setting: Attacker can query the model and get the predicted label



- Example: CIFAR10 image classification
  $y \in \{1, \cdots, K\}$, K: total number of classes

Substitute Attack

- Intuition: adversarial examples transfer across models

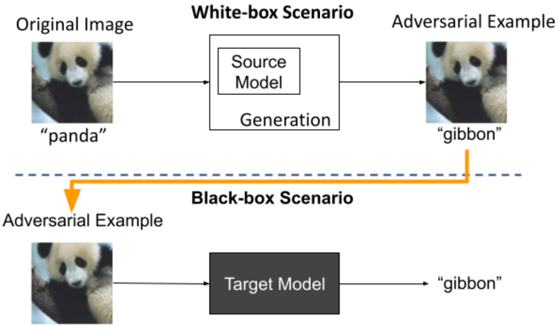
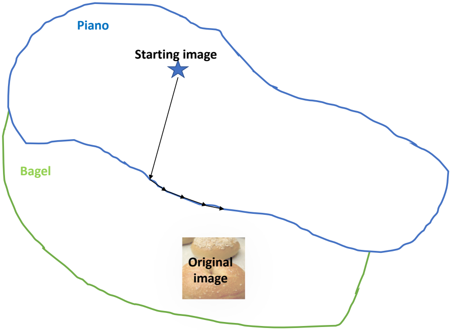


Figure from link

- Train a substitute model using a small amount of training data.
- Generate adversarial examples based on the substitute model. (Papernot et al., 2017)

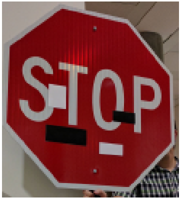
Boundary Attack



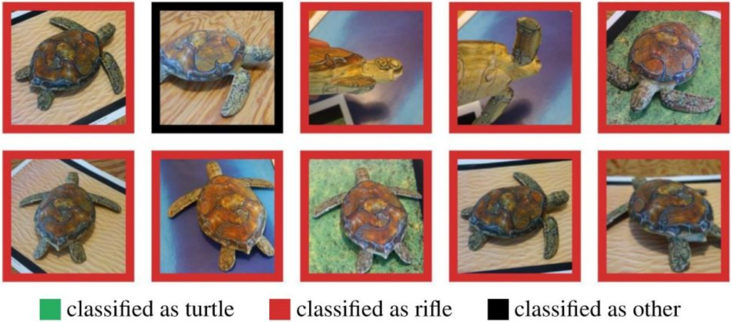
- Start as a random example from a target class.
- Perform rejection sampling along the boundary.

  (Brendel et al., 2018)

Attack in Physical World



Attack in Physical World



Conclusions

- Adversarial examples
- Different types of adversarial attacks

Questions?