INT307 Multimedia Security System

Generative Learning

Shengchen Li

Xi'an Jiaotong-Liverpool University

12th Nov 2022

Aims

- Understand basic knowledge related to adversarial attacks of deep learning systems
- Know the concept of algorithm robustness of deep learning systems

Adversarial Attack

- Modify the pictures to mislead machine learning algorithms

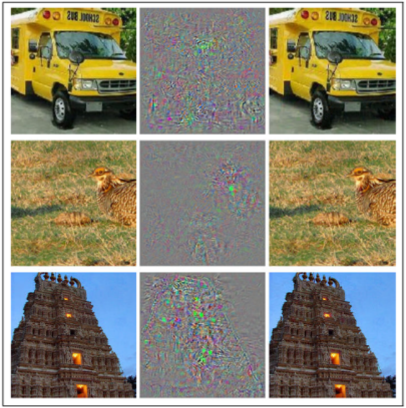

Figure 1-2. Subtle perturbations result in image misclassification—the original images are on the left, and the perturbed images on the right were all misclassified as "ostrich" (image from Szegedy et al. 2014)

Adversarial Perturbation

- Altering data samples by a tiny amount to mislead the machine learning algorithm

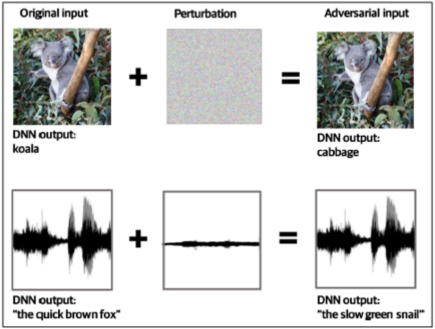

Figure 1-3. Adversarial perturbation applied across an image to fool an image classifier and across audio to fool a speech-to-text system

Unnatural Adversarial Input

- Sometimes, the diagrams are not even similar with the label

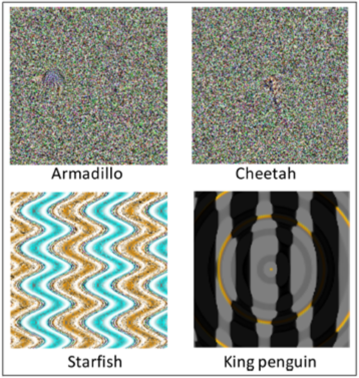

Figure 1-4. Digitally generated adversarial examples that are unrecognizable by humans with associated classifications from state-of-the-art DNNs (image from Nguyen et al. 2015)

Adversarial Patches

- Change a small area to cheat the classifier (Maximise Diversion)

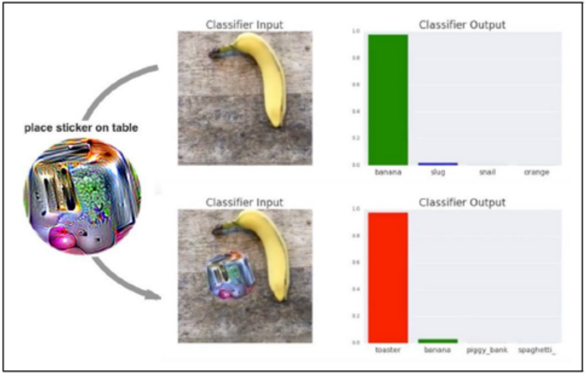

Figure 1-5. Digitally generated adversarial patch causing confident misclassification of a banana as a toaster (from Brown et al. 2017[6])

Attacks to Deep Learning systems

- Targeted attacks: Cause DL systems return an incorrect result
- Untargeted attacks: Cause DL systems return an expected wrong output

Recall: Feature Space

- Deep Neural Network projects raw media to a feature space

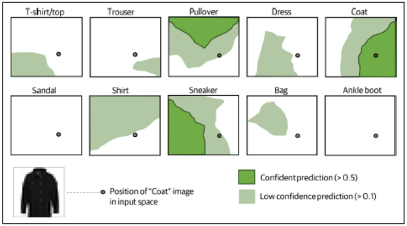

Figure 5-2. A model's prediction landscapes for each classification—zoomed into a tiny area of the complete input space

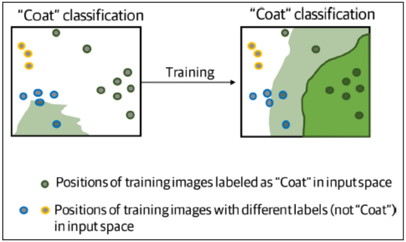

Figure 5-3. The changing prediction landscape of the input space during training

OoD Data

- Out of Distribution data does not follow the distribution of the feature space (out of domain data)

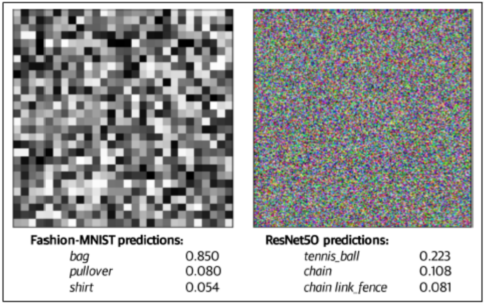

Figure 5-4. Classification predictions for randomly generated images

Perturbation Attack

- The principle of perturbation attack is to use minimum change to cause maximum impact

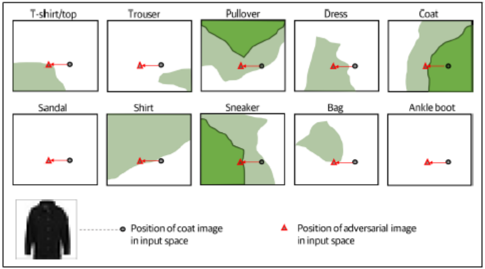

Figure 5-8. Untargeted attack—moving outside the "Coat" classification area of the input space

White Box Methods

- JSMA: Jacobian Saliency Map Approach
- Calculate adversarial saliency of each input
  - The effect of the change on increasing the predicted score for the target classification (in a targeted attack)
  - The effect of the change on decreasing the predicted score for all other classifications

Limited Black Box Methods

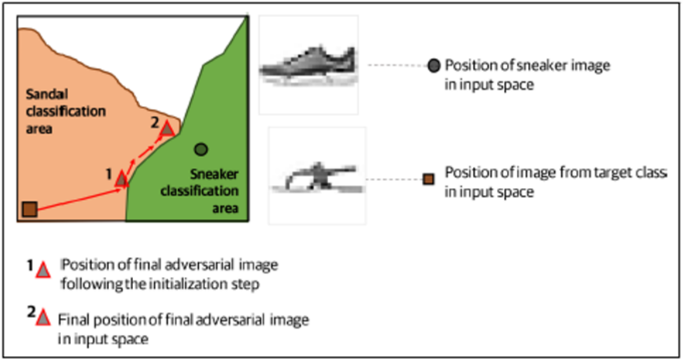

Figure 6-9. Targeted limited black box boundary attack

We usually have a target class for the attack

Boundary Attack

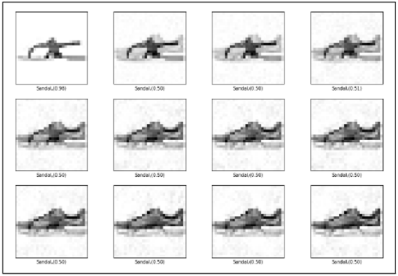

Figure 6-13. The boundary attack begins with an image from the target class and gradually moves it closer to the original

Score-based Black Box Methods

- Usually based on prior knowledge as well

- One way is to analyse the resulting score predicted

Attack Modes

- Direct attack: The attacker develops the attack on the target system itself
- Replica attack: The attacker has access to an exact replica of the target DNN in order to develop the attack
- Transfer attack: The attacker develops the attack on a substitute model which approximate the target
- Universal transfer attack: The attacker has no information about the target model. They create adversarial input that works across an ensemble of models that perform similar functions to the target in the hope that it will also work on the target DNN

Physical-World Attacks

Difficulties:

- Creations of the adversarial input
- Capture of the adversarial input
- Effects of positioning and proximity of adversarial input with respect to the sensor
- Environmental conditions
- Attack constrains

Adversarial Objects

- Object Fabrication and Camera Capabilities
  - 3D or 2D printing
- Viewing Angles and Environment
  - Viewing (Zoom, Rotation, Skew)
  - Lighting

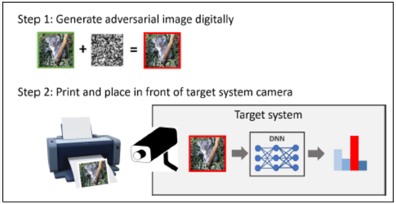

Figure 8-1. Highly constrained physical-world adversarial attack using a printed image

Adversarial Sound

- Audio Reproduction
- Microphone Capabilities
- Audio Positioning
- Environment

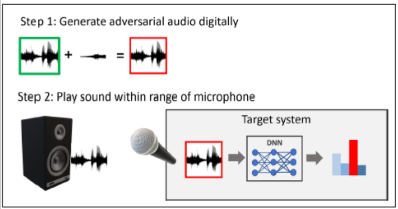

Figure 8-6. Physical-world adversarial attack using sound

Aim of Adversarial Attack

- Capabilities
- Ability to affect the input
- Knowledge of access to the target

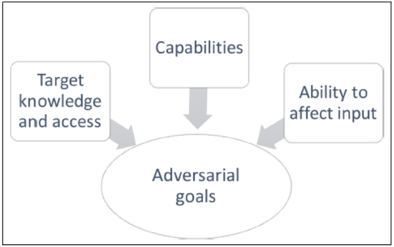

Figure 9-4. Adversarial goals are constrained by target knowledge and access, adversarial capability, and the attacker's ability to affect the input.

Model Evaluation: empirically studies

- **Difficulties on Measurement**
  - Hard to predict the features of adversarial data
  - Low likelihood of reduplicate adversarial data

- **Possible Attempt**
  - Threat model
  - Attack Methodology
  - Test Data

Example

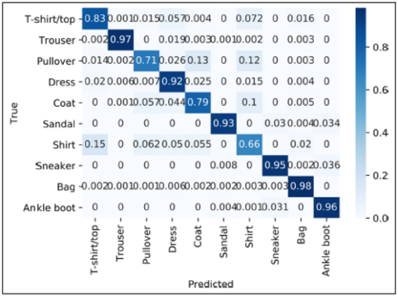

Figure 9-5. A confusion matrix for the Fashion-MNIST classifier provides a summary of the model's performance for each of the fashion labels.

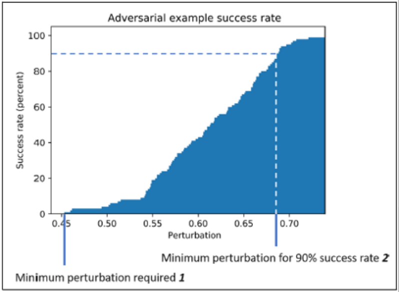

Figure 9-6. Allowing greater perturbation increases the success rate for an adversarial example.

Theoretically Derived Robustness Metrics

- Measures the safe zone to adversarial attack in a feature space

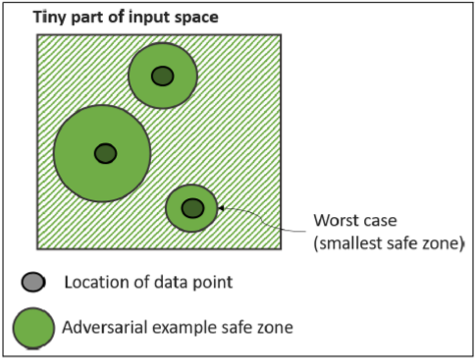

Figure 9-7. Calculating adversarial example safe zones in the input space

Theoretically Derived Robustness Metrics

1. Improve the model

2. Remove adversarial aspects from input

3. Minimise the adversary's knowledge

Theoretically Derived Robustness Metrics

1. Gradient masking: Knowledge Distillation
2. Adversarial training
3. Out-of-distribution (OoD) detection
4. Randomised dropout uncertainty measurements