INT 307
Multimedia Security System

Neural Network and Adversarial Attack II

Sichen.Liu@xjtlu.edu.cn

XJTLU | SCHOOL OF FILM AND TV ARTS

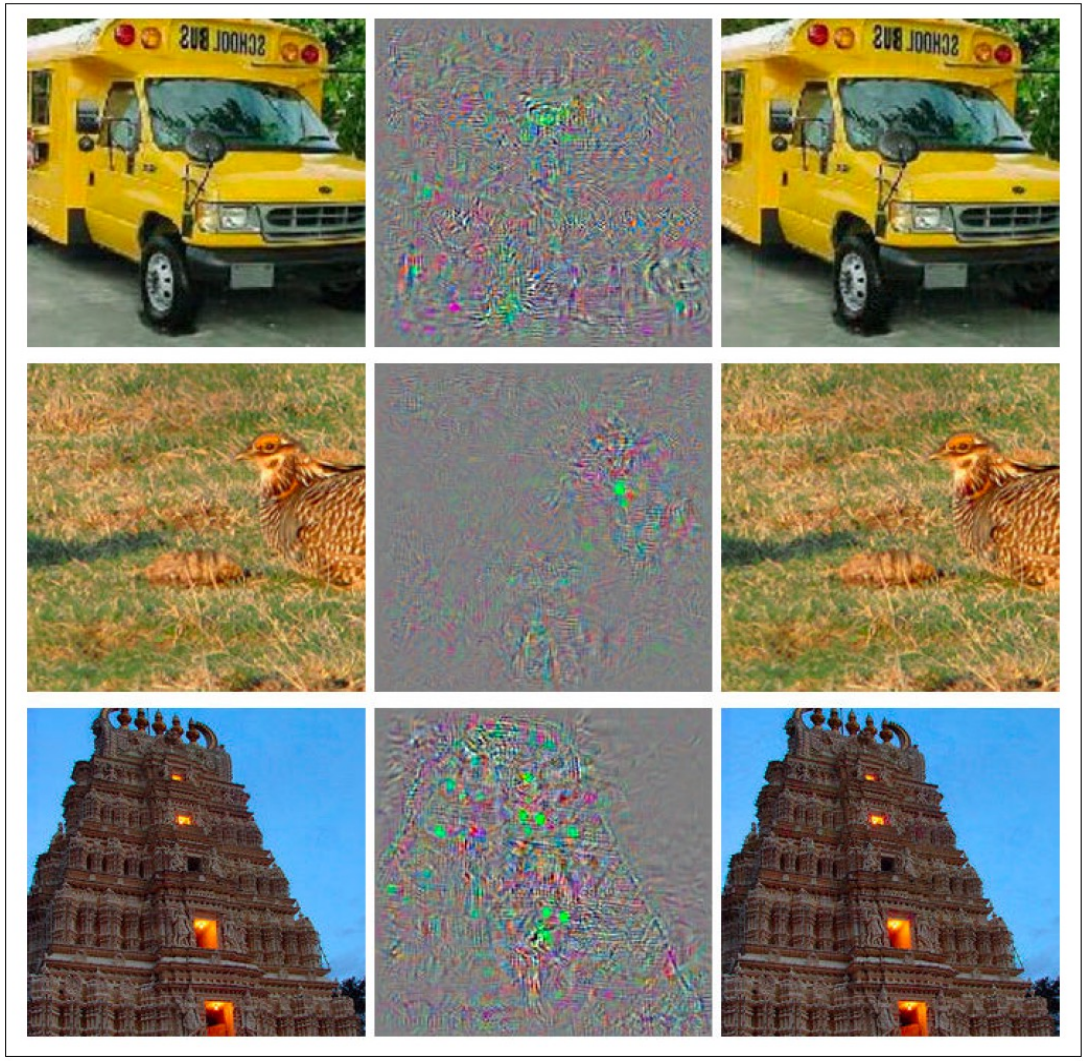
Xi'an Jiaotong-Liverpool University
西交利物浦大學

Aims

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

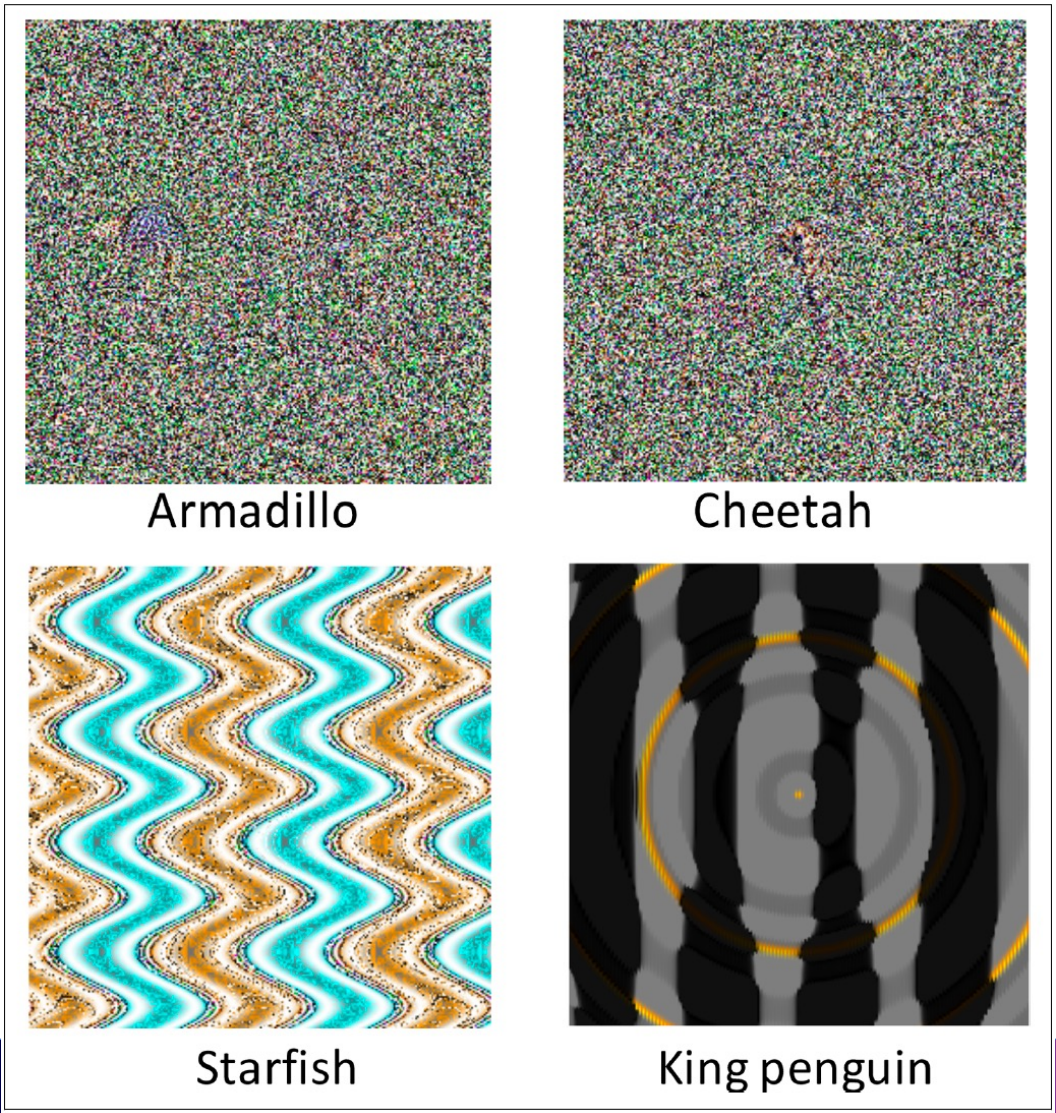
Adversarial Attack

- Modify the pictures to mislead machine learning algorithms



Unnatural Adversarial Input

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



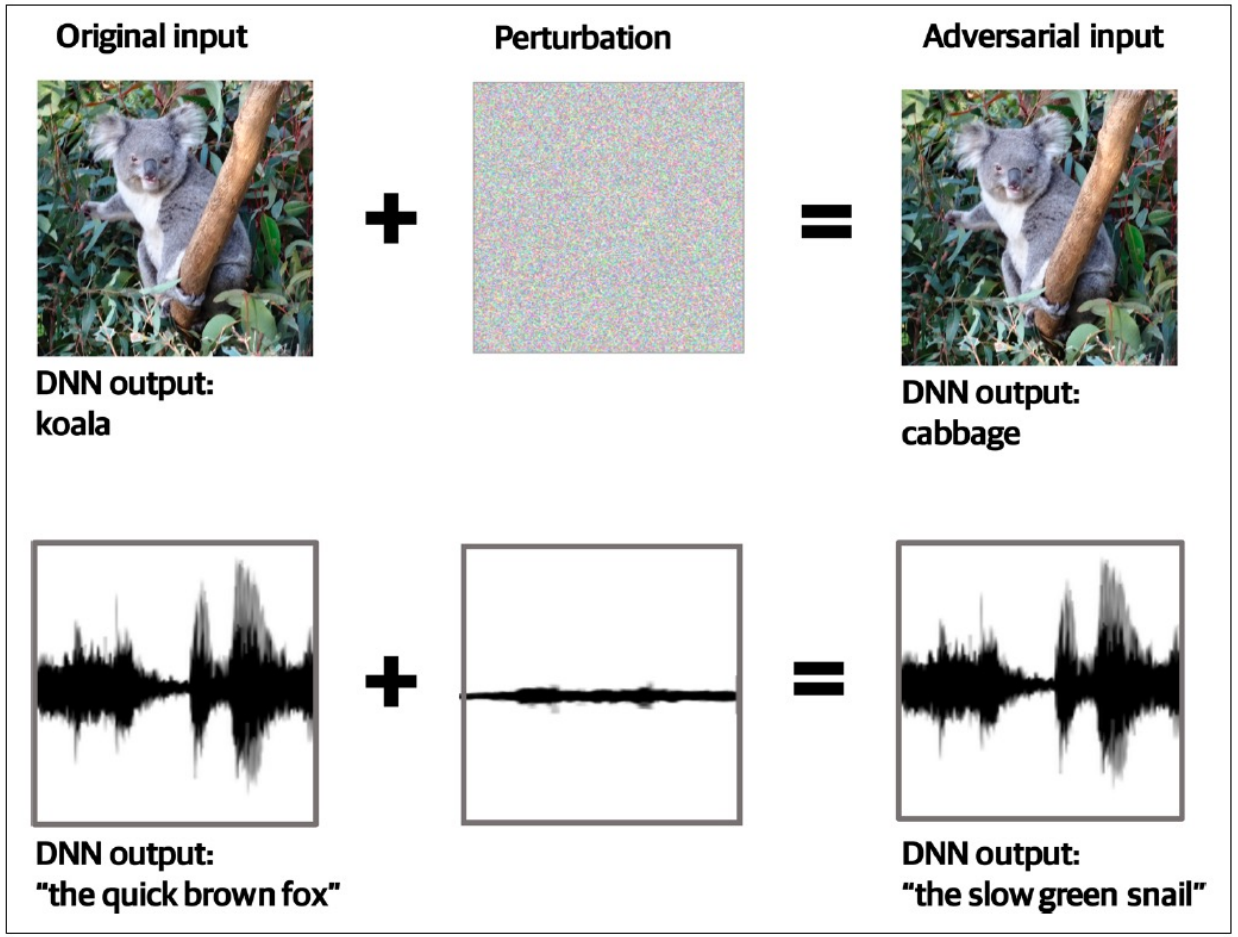
Adversarial Perturbation & Adversarial Patches

- **Adversarial perturbation**
  A combination of imperceptible (or nearly imperceptible) small changes distributed across the input data which cause the model to return an incorrect result.

- **Adversarial patch**
  An addition to a specific area (spatial or temporal) of the input data to cause the model to return an incorrect result. An adversarial patch is likely to be perceptible by a human observer, but could be disguised as something good.

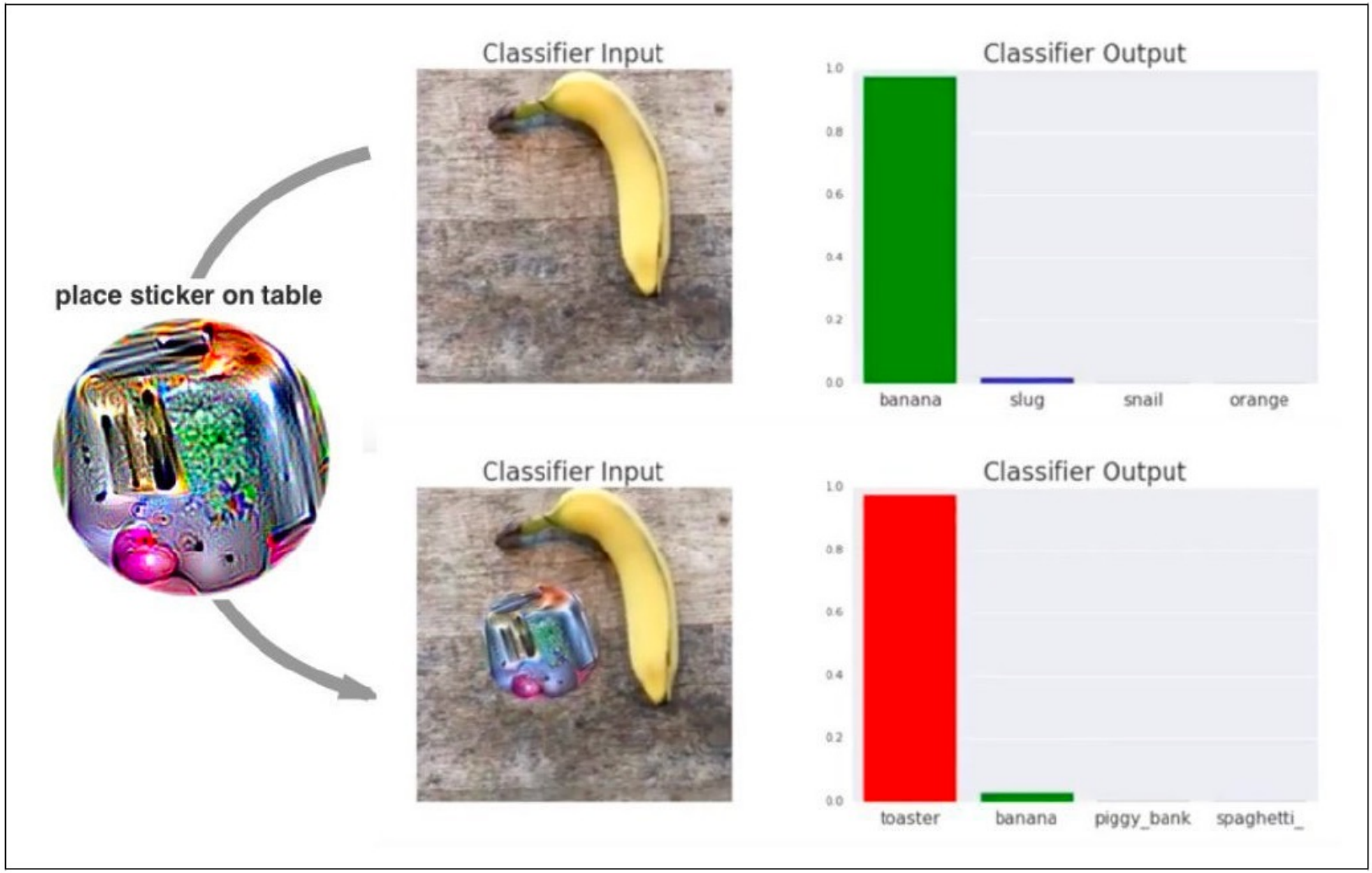
Adversarial Perturbation

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



Adversarial Patches

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



Attacks to Deep Learning systems

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

Feature Space

- Deep Neural Network projects raw media to a feature space

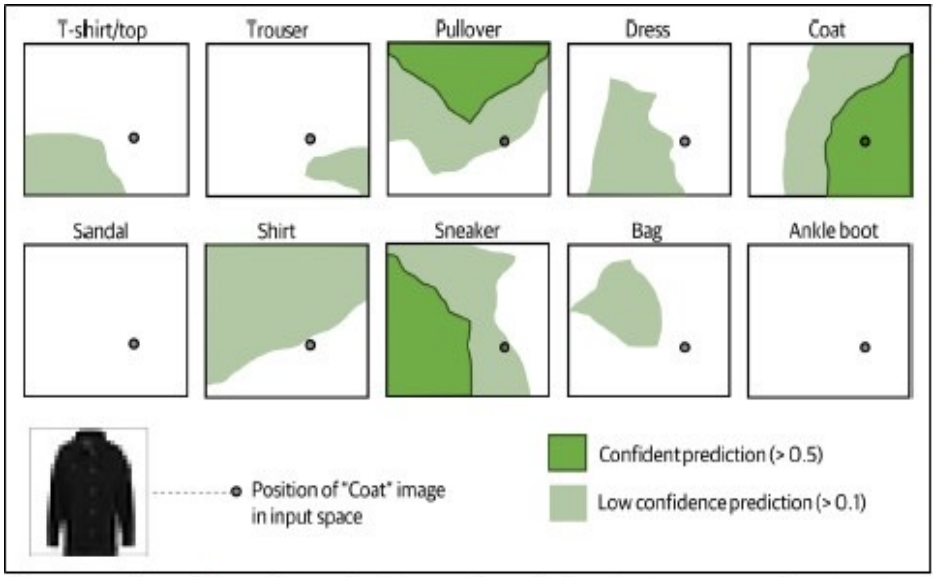


Figure 5-2. A model's prediction landscapes for each classification—zoomed into a tiny area of the complete input space

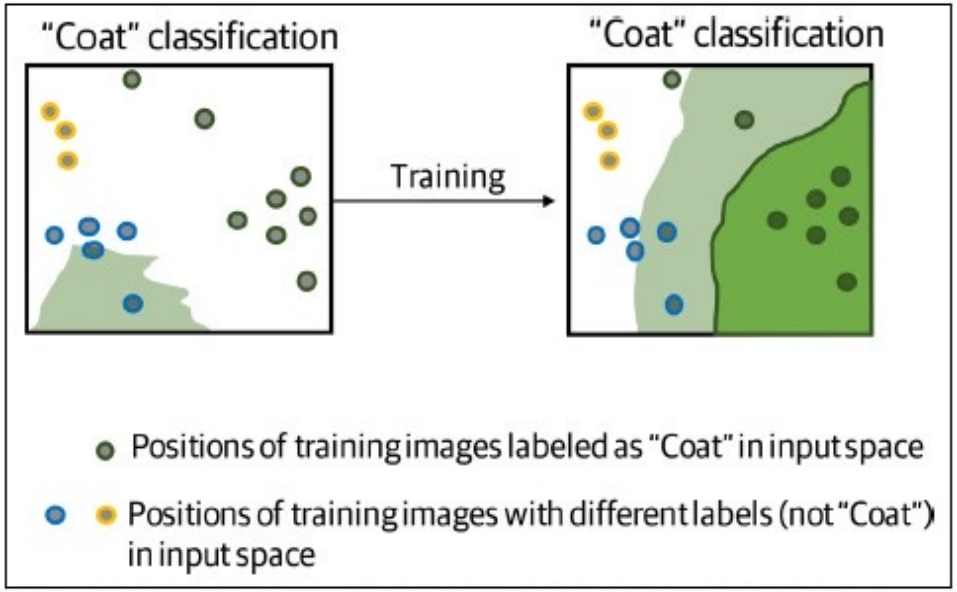
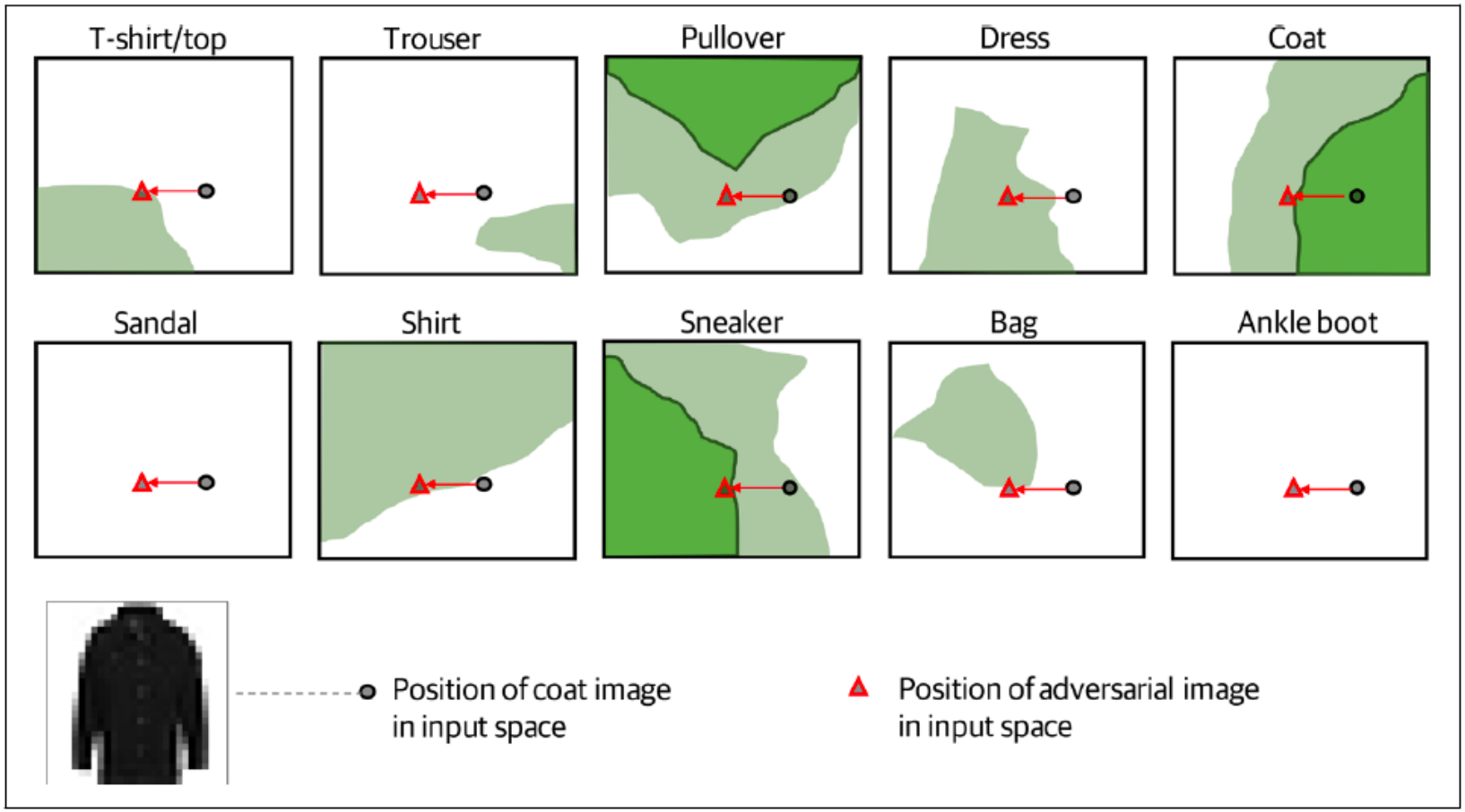


Figure 5-3. The changing prediction landscape of the input space during training

Perturbation Attack

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



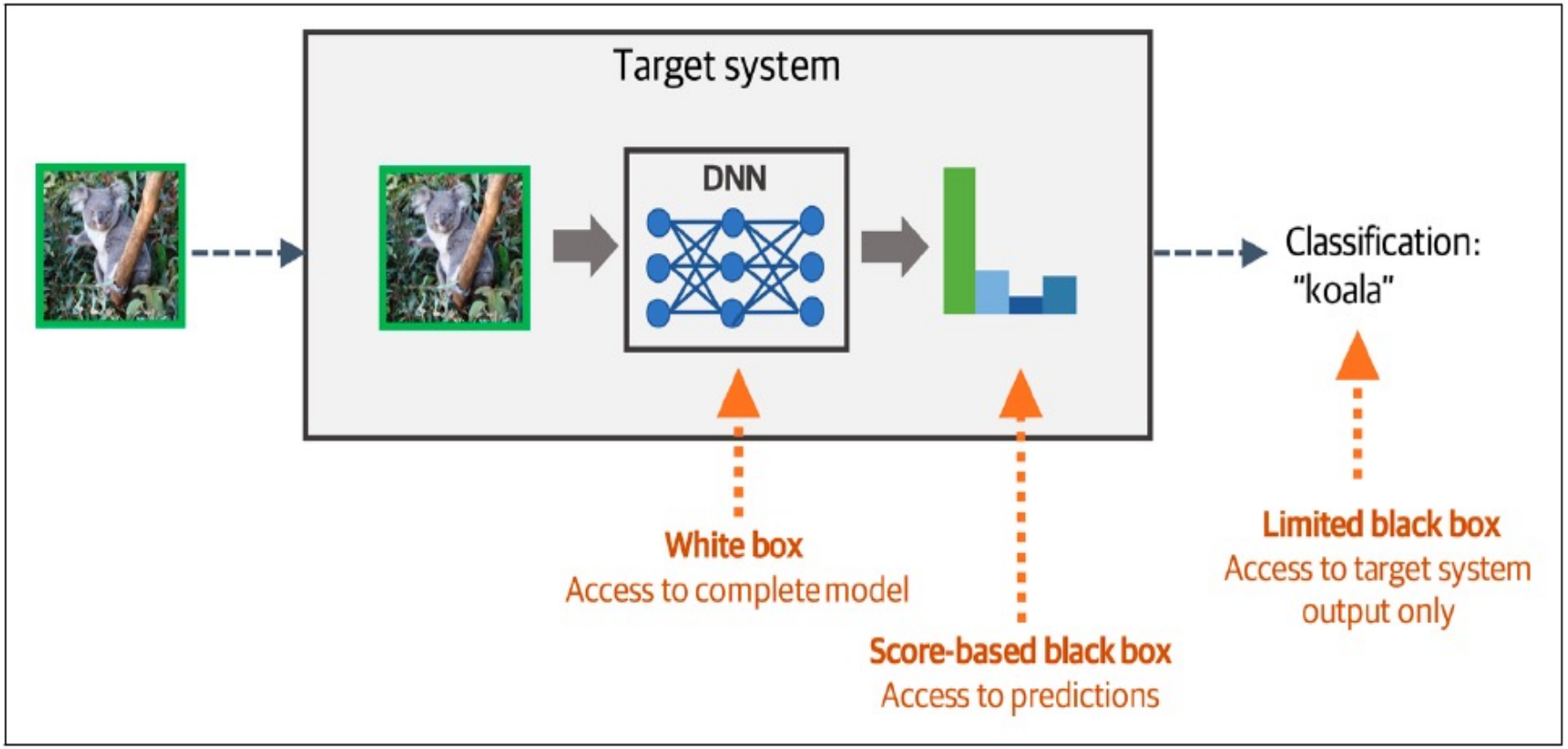
Methods for Generating Adversarial Perturbation

- **White box**
  These methods exploit complete knowledge of the DNN model to create adversarial input.

- **Limited black box**
  These methods refine adversarial input based on an output generated from the model or from the system in which it resides. For example, the output might be simply a final classification.

- **Score-based black box**
  These methods refine adversarial input based on the raw predictions (scores) returned from the DNN. Score-based methods lie somewhere between white box and limited black box methods.

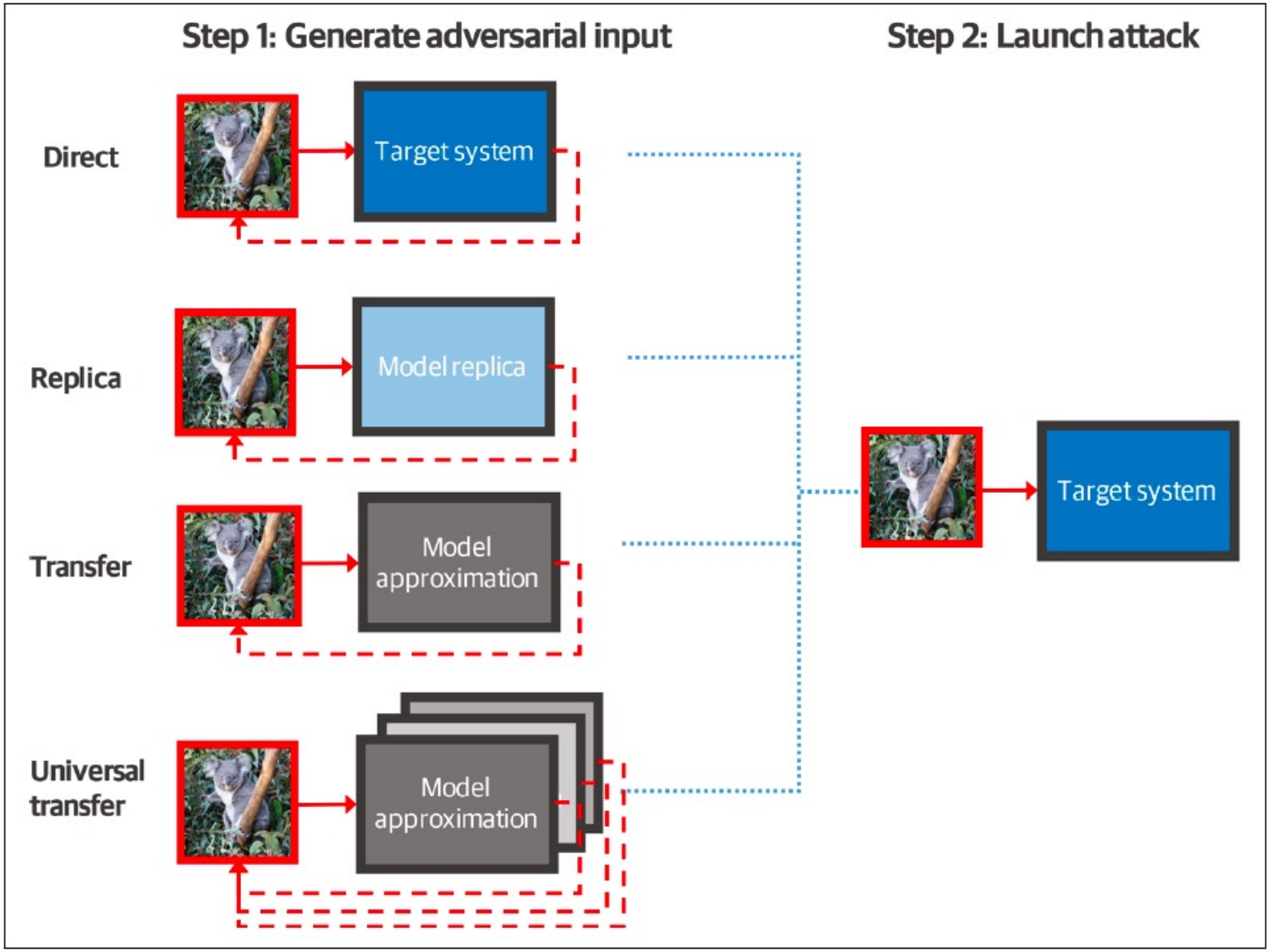
Methods for Generating Adversarial Perturbation



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

Attack Modes



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

Adversarial Objects

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

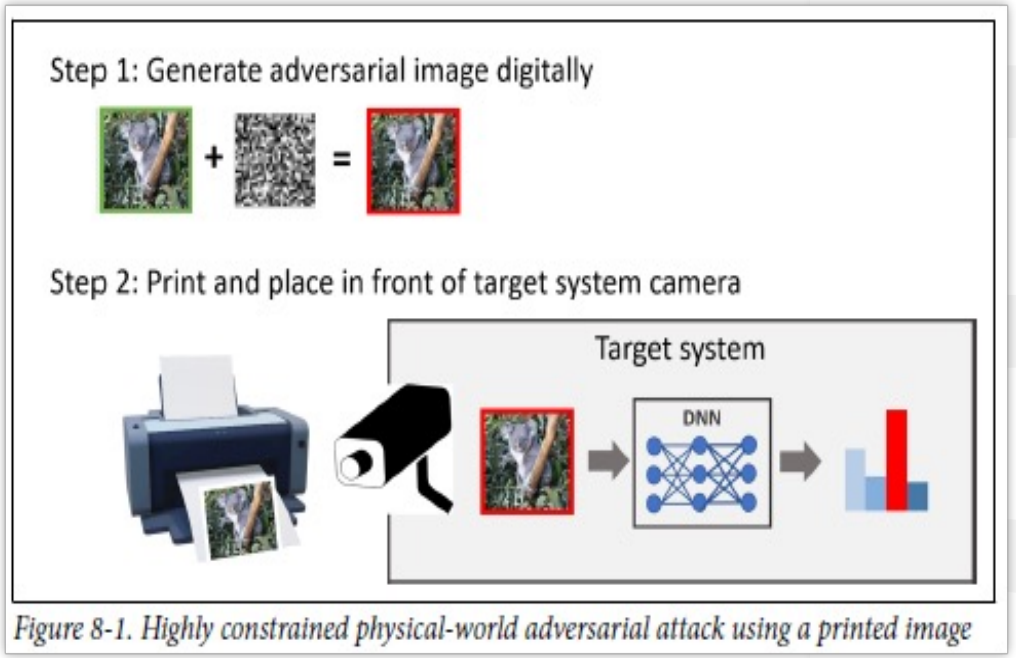


Figure 8-1. Highly constrained physical-world adversarial attack using a printed image

Adversarial Sound

- Audio Reproduction
- Microphone Capabilities
- Audio Positioning
- Environment

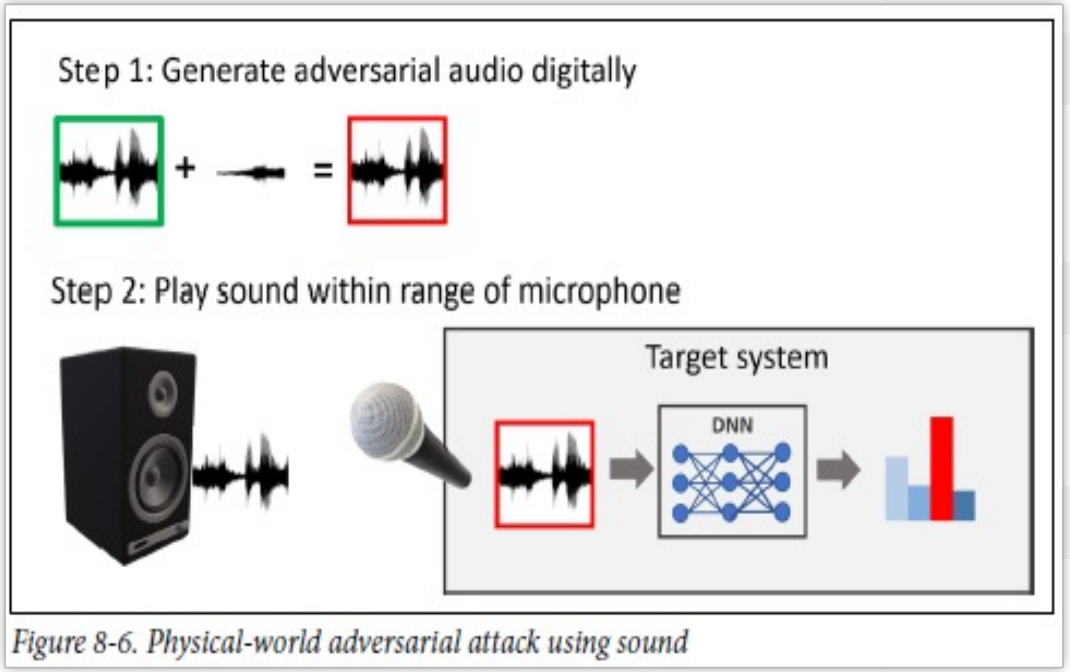


Figure 8-6. Physical-world adversarial attack using sound

Theoretically Derived Robustness Metrics

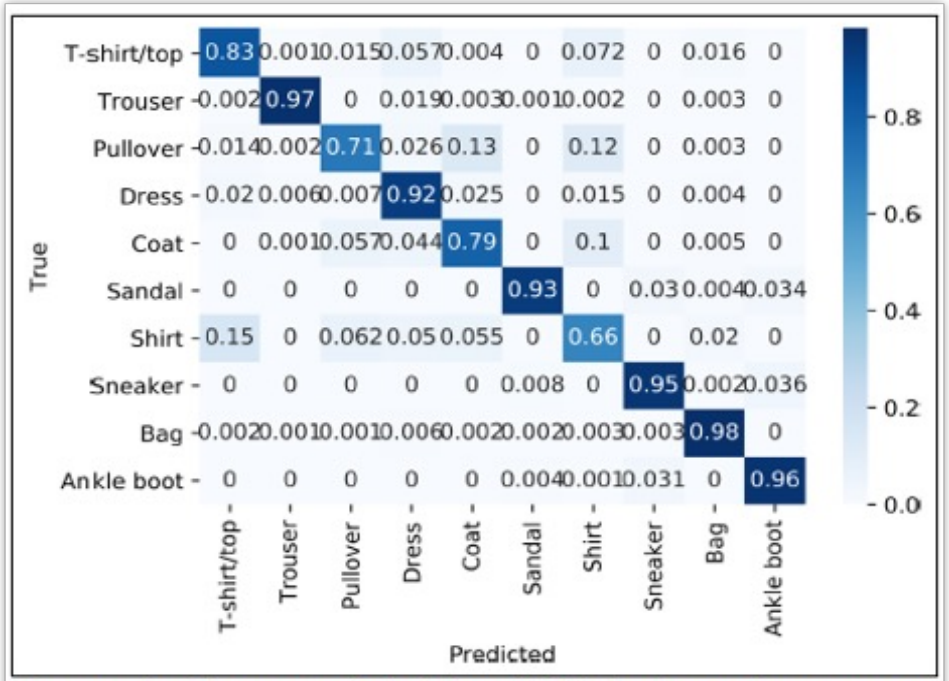


Figure 9-5. A confusion matrix for the Fashion-MNIST classifier provides a summary of the model's performance for each of the fashion labels.

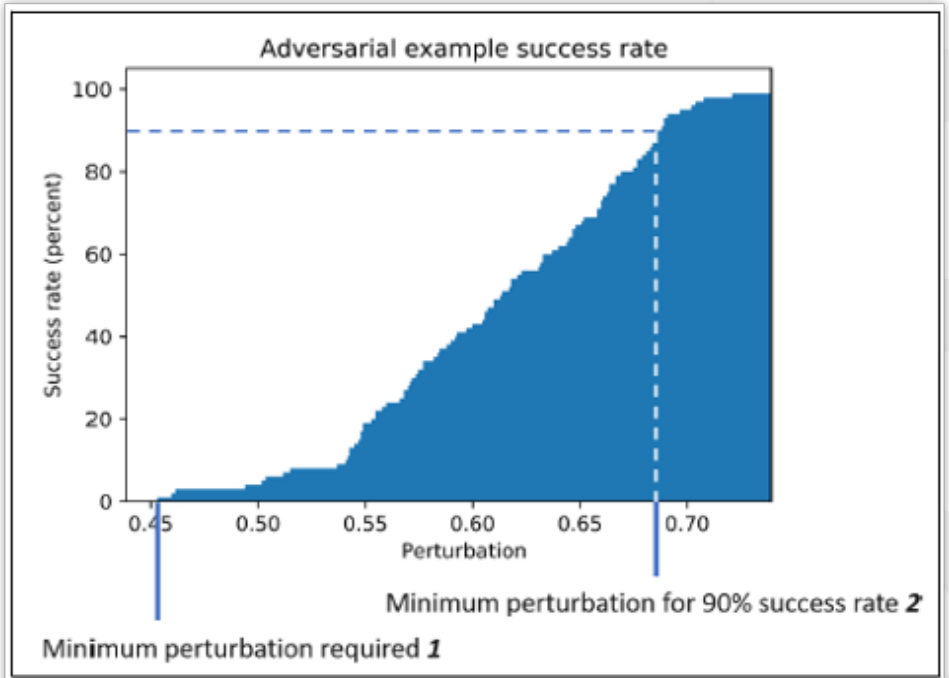


Figure 9-6. Allowing greater perturbation increases the success rate for an adversarial example.

Theoretically Derived Robustness Metrics

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

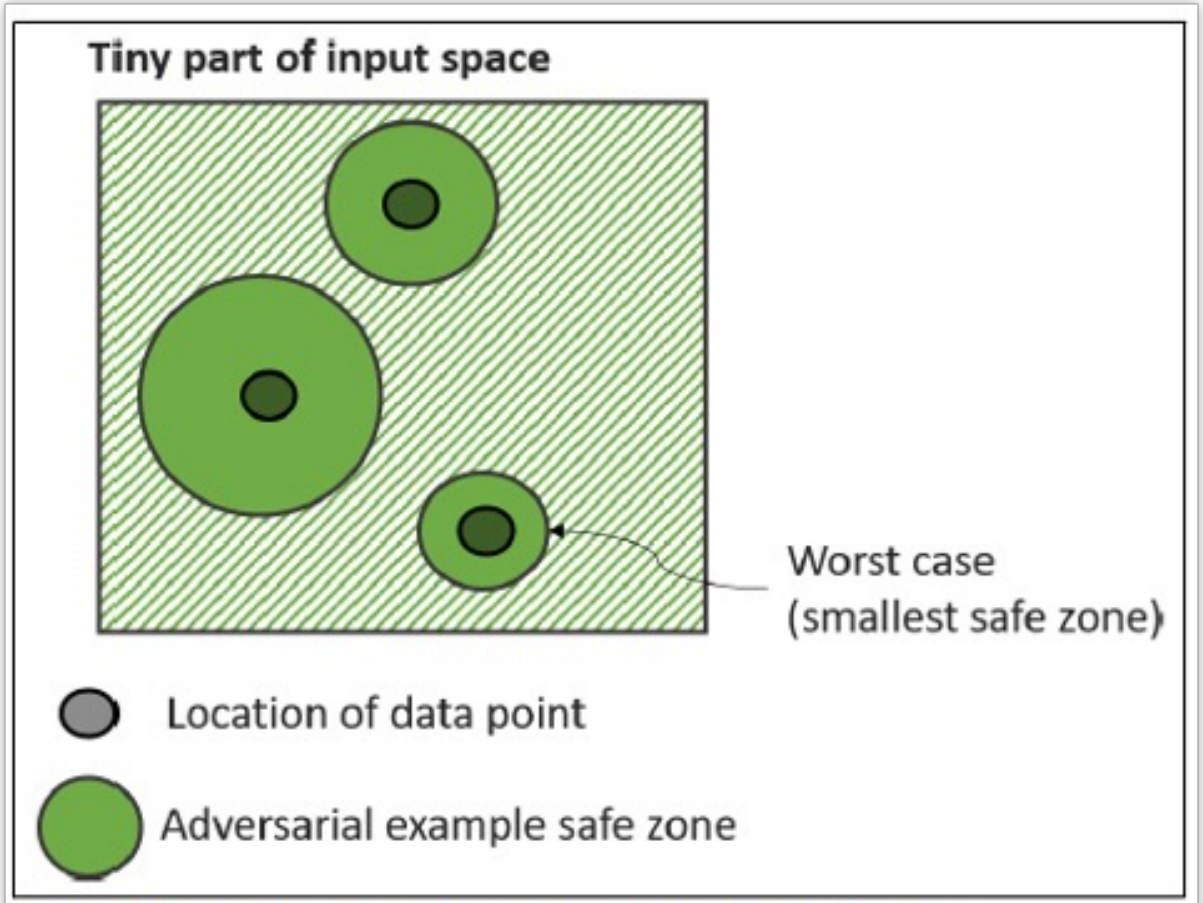
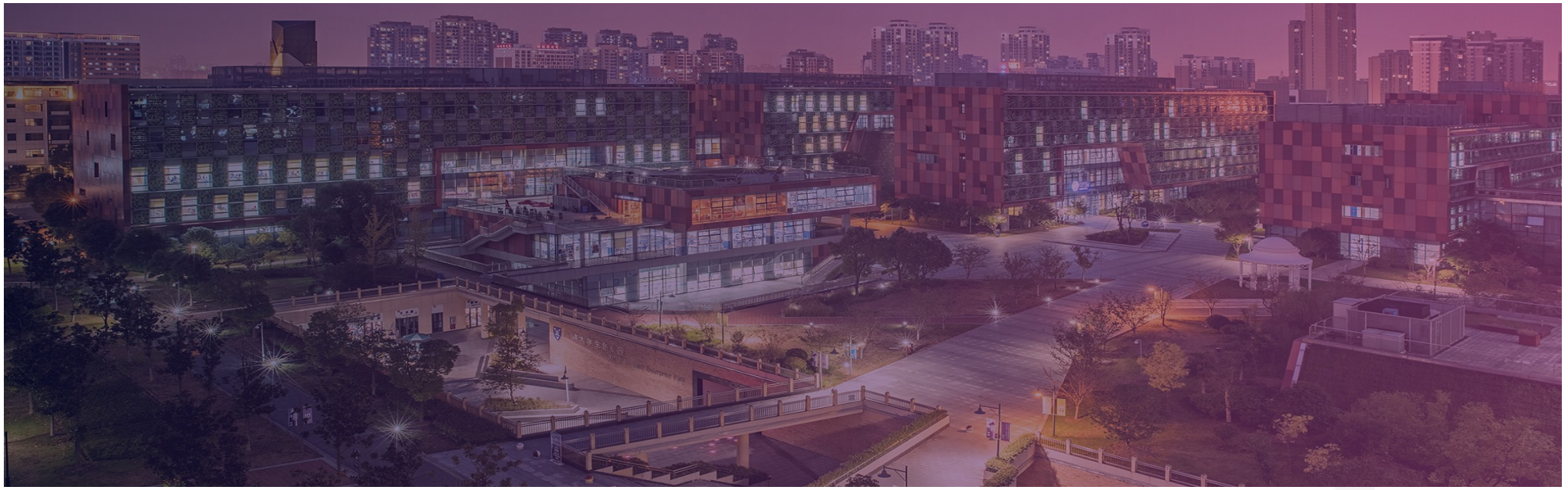


Figure 9-7. Calculating adversarial example safe zones in the input space

Theoretically Derived Robustness Metrics

1. Adversarial training
2. Defensive Distillation
3. Feature Squeezing
4. Input Transformation
5. Randomised dropout uncertainty measurements
6. Minimise the adversary's knowledge



THANK YOU

VISIT US

WWW.XJTLU.EDU.CN

FOLLOW US

@XJTLU

Xi'an Jiaotong-Liverpool University
西交利物浦大学

XJTLU | SCHOOL OF FILM AND TV ARTS