Adversarial machine learning

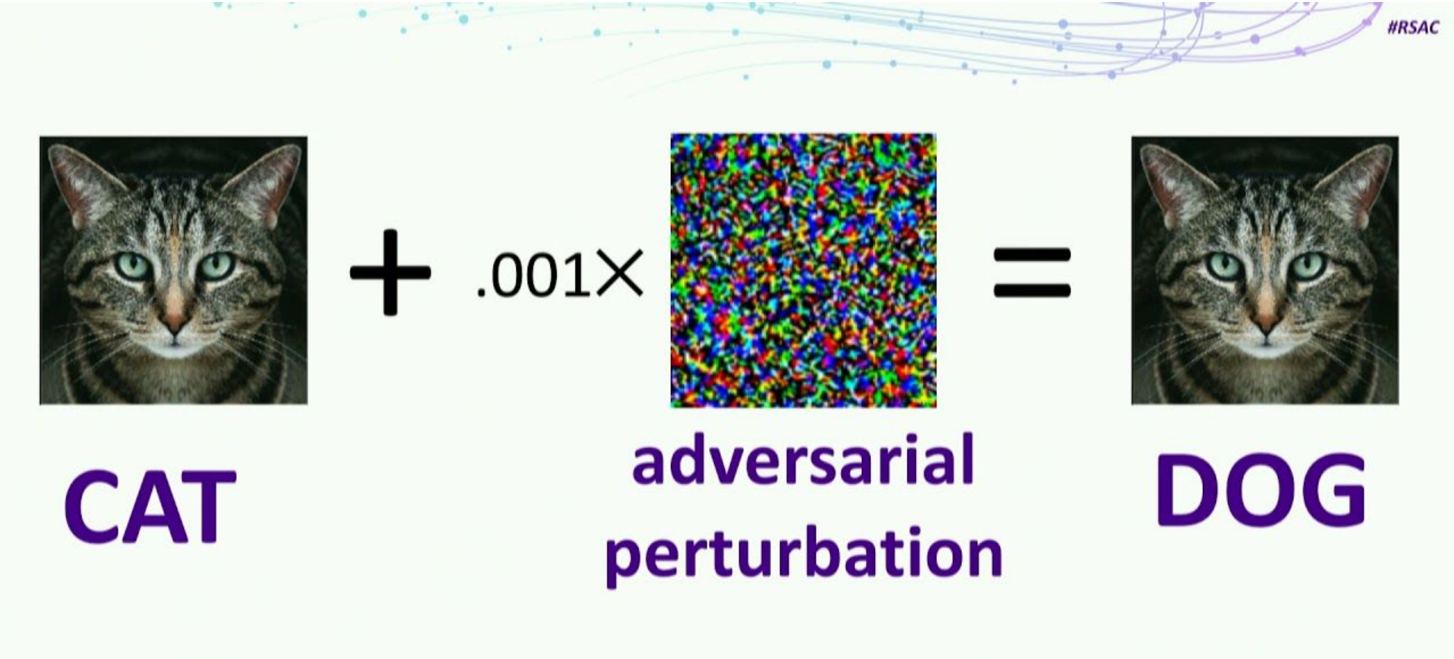
Proposal presentation

Nathan Decou
Killian Trolès

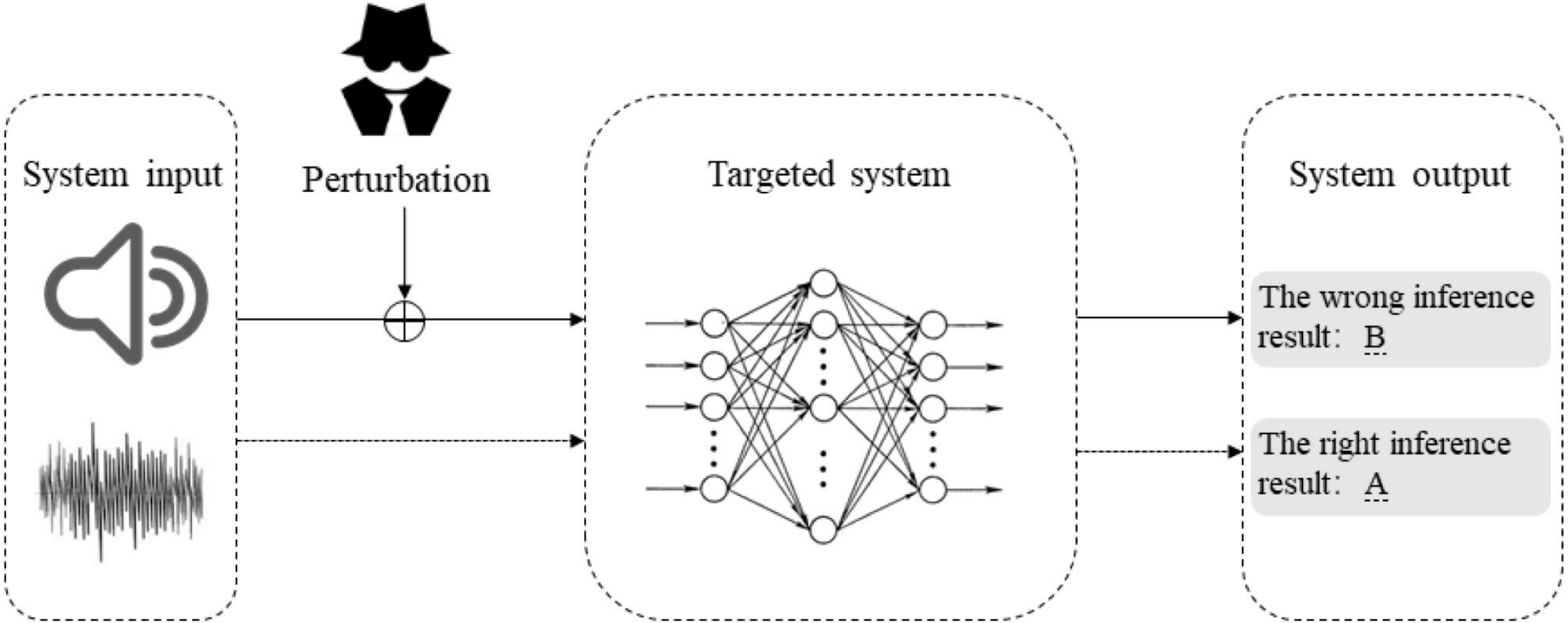
ILLINOIS INSTITUTE OF TECHNOLOGY

“Adversarial machine learning is a machine learning technique that attempts to fool models by supplying deceptive input.”

Wikipedia



Source: I.J Goodfellow, J. Shlens and C. Szegedy. Explaining and harnessing adversarial examples. 2015



Source: Xiaojiao Chen ,Sheng Li and Hao Huang. Adversarial Attack and Defense on Deep Neural Network-Based Voice Processing Systems: An Overview

Use of image recognition

- Autonomous vehicle
- Military surveillance
- Person identification
- Medical imagery - disease diagnosis

Goals of this project

- Demonstrate that an image recognition model can be misled by minor modification to its input
- If time, develop a model resistant to this kind of attacks

How we will do it

- Find methods to create the right mask to apply to trick the model
- Add noise[1] to images using those methods so that
  - the model cannot associate them the right label
  - the model associate them another label, previously chosen

[1] noise : minor modification, invisible to human eyes

Evaluation methods

- Confusion Matrix
- Model confidence after perturbation
- Similarity metrics to measure the similarity between original and adversarial images

Any question?