Impact of AI Decal:
Optical Illusions for Neural Networks

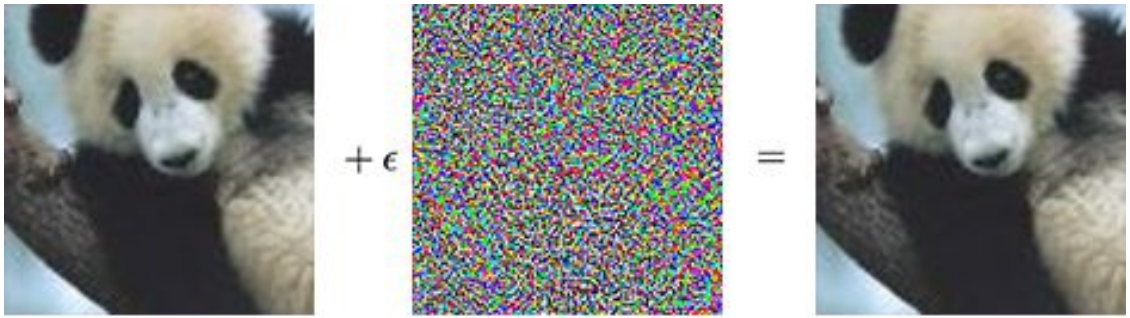
Gokul Swamy & Brenton Chu

Quiz: https://tinyurl.com/impactsp19q6

Interpretability in Machine Learning

Adversarial Examples

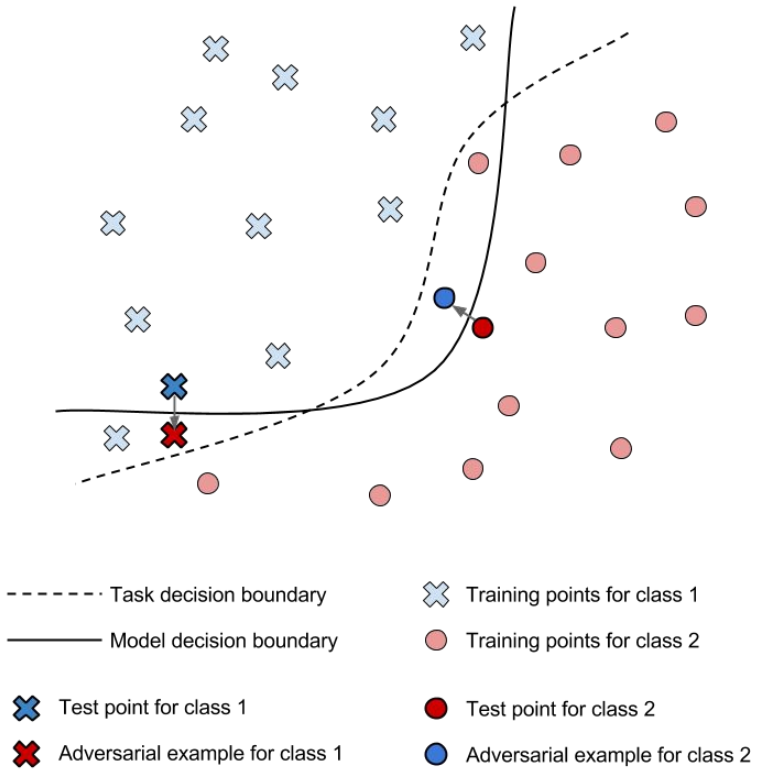
➔ “Inputs to machine learning models that an attacker has intentionally designed to cause the model to make a mistake; they’re like optical illusions for machines” - OpenAI



Adversarial Examples

➔ Not just limited to digital inputs, can also happen in real world

How do adversarial examples work?

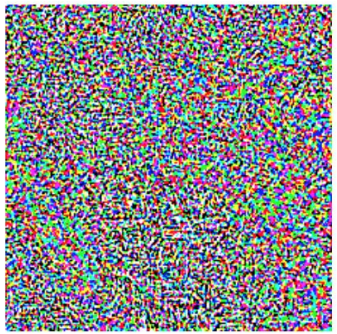


How do adversarial examples work?

$\boldsymbol{x}$

"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

Adversarial Examples that also fool people

Adversarial Examples that also fool people

Yanny or Laurel?

➔ https://www.nytimes.com/interactive/2018/05/16/upshot/audio-clip-yanny-laurel-debate.html

Q:

What are some consequences of AI that can be fooled this way?

Consequences

Lab (Stationary) Test

Physical road signs with adversarial perturbation under different conditions

Cropping, Resizing

Stop Sign → Speed Limit Sign

Field (Drive-By) Test

Video sequences taken under different driving speeds

Sample Per K Frames, Cropping, Resizing

Stop Sign → Speed Limit Sign

Consequences

https://nicholas.carlini.com/code/audio_adversarial_examples/

Impact of AI Decal: *Activity*

Activity

- ➔ We’re going to have a socratic seminar on set of discussion prompts
- ➔ Rules for discussion:
  - ◆ Please don’t dominate the discussion
  - ◆ If someone isn’t talking much, bring them into the conversation
  - ◆ Ask clarifying questions
  - ◆ Support your points with evidence or logic
  - ◆ You are free to disagree, but please do so courteously
  - ◆ Be kind to one another

Socratic Seminar

- ➔ Are adversarial examples a problem worth working on?
- ➔ How can we protect ourselves on overdependence on fallible AI?
  - ◆ Is Human-in-the-Loop enough?
- ➔ Which is safer, a highly accurate but uninterpretable system or a less accurate but interpretable system?
- ➔ What are some current systems used today that could be susceptible to adversarial examples, and harmful would fooling those systems be?

Writing Assignment Two

- ➔ Same process as for the previous writing assignment
- ➔ Address one of these prompts:
  - ◆ Briefly describe a best-case world with self-driving cars and nearly complete automation. What are some of the biggest hurdles that we need to overcome to achieve such a world? Of those hurdles, pick the one you think is most important and explain it in detail.
  - ◆ What are the most dangerous examples of adversarial examples, and what are the consequences if they are left unaccounted for? For the examples you have mentioned, what are the most effective strategies (technological, policy, or otherwise) to prevent or mitigate potential harm?
  - ◆ Take any topic/issue that has been discussed in class up to this point and elaborate on it. What is a potential solution, and why would it work? Alternatively, what is a currently proposed solution that you think may not work, and why?

Impact of AI Decal:

Next: Artificially Generated Data