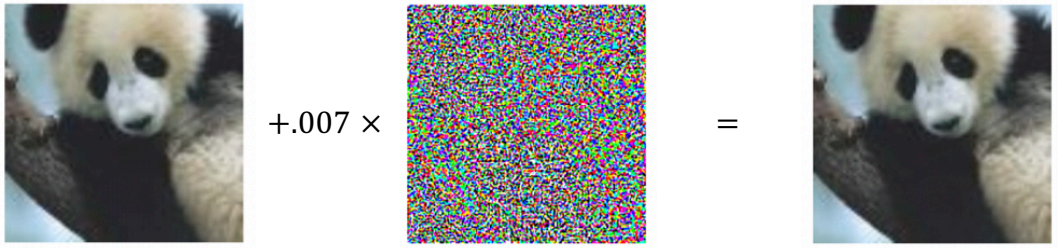
When AI Breaks

Introduction

- Only slight alterations to an input image (in the context of image classification) can drastically fool a deep learning model which would otherwise classify an image correctly (e.g., panda) into outputting a completely wrong label (e.g., gibbon).



- This phenomenon occurs even when the perturbations are imperceptible to the human eye.
- These "doctored" images are called **adversarial examples**. Making neural networks robust to these attacks is an increasingly active ML research area.

DataScience@SMU

Not Just DL

Not Unique to DL

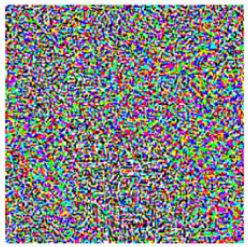
1. It's important to note that, while these techniques are not unique to DL, they are exacerbated. Even an RF algorithm trained on the iris dataset will make predictions incorrectly but confidently if the provided sample is vastly different (i.e., 3-foot by 3-foot flower).
2. There are two reasons this is so concerning in DL:
    1. We understand so little of the structure that the algorithm has learned.
    2. The algorithm learns from the data so closely to gain an increased accuracy that it is more susceptible to overfitting on the provided distribution.

DataScience@SMU

Case Study 1

Panda vs. Gibbon: Case Study

$+.007 \times$



$=$

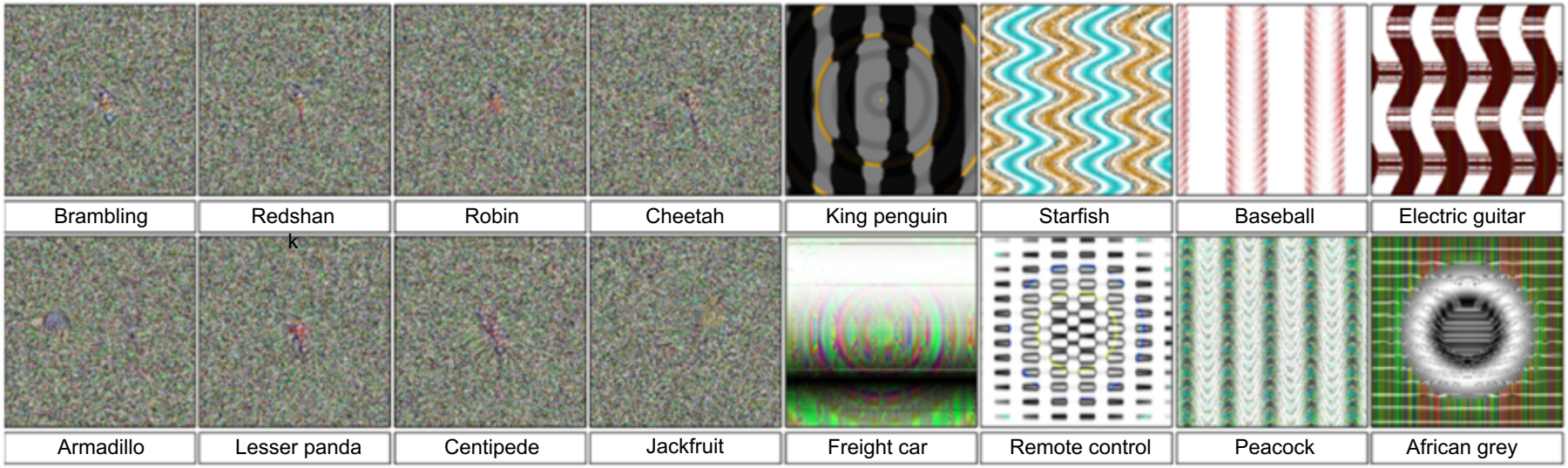
x
"Panda"
57.7% confidence

$\text{sign}(\nabla_x J(\theta, x, y))$
"Nematode"
8.2% confidence

$x +$
$\epsilon\text{sign}(\nabla_x J(\theta, x, y))$
"Gibbon"
99.3 % confidence

Direct Encoding

Indirect Encoding



Adversarial Examples: *Attacks and Defenses for Deep Learning* (https://arxiv.org/pdf/1712.07107.pdf)

Seeing Ghosts

1. “Cloud face” is a 2013 installation showing the ”faces” that face recognition algorithms identify in clouds.
2. Deep Dream is a 2014 DL “hallucination” of what algorithms see when they analyze images.

http://ssbkyh.com/works/cloud_face/

Other Adversarial Examples

- Fake planes were developed to confused WWII bombers into dropping bombs in the wrong places.
- As the story goes, “cargo cults” were supposedly developed in the Pacific arena where wooden planes were sufficient to trick high-flying resupply missions to drop their supplies to the locals rather than the intended troops.
- Silicon face masks have been used to confuse border crossing agents for political asylum seekers.
- In each case, the adversarial exploit relied upon a miss classification. Deep learning promises deep insights by developing complex models, but inherently such models are more brittle then less complex ones.

Case Study

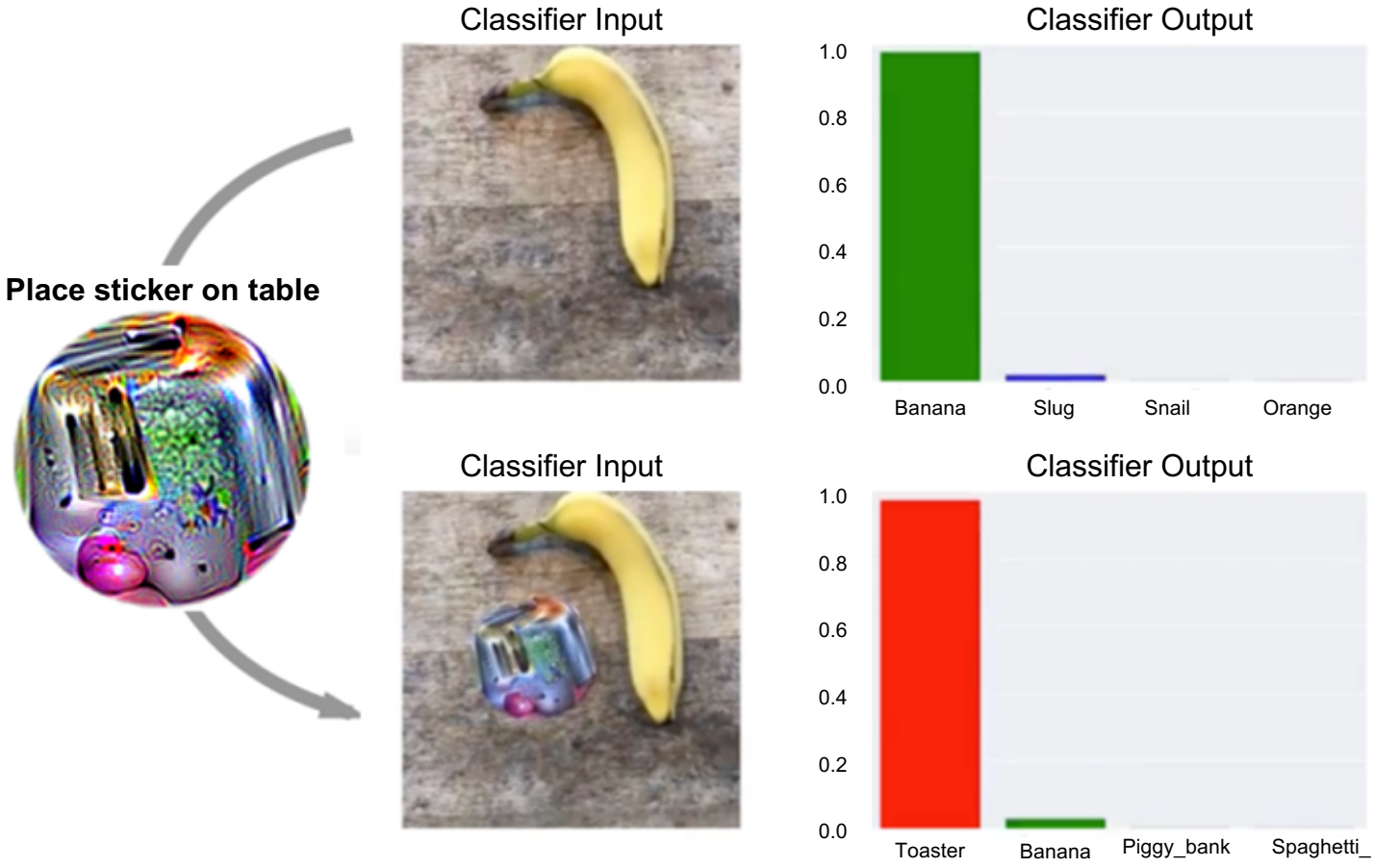
(a) (b) (c) (d)

Accessorize to a Crime: Real and Stealthy Attacks on State-of-the-Art Face Recognition
https://www.cs.cmu.edu/~sbhagava/papers/face-rec-ccs16.pdf

DataScience@SMU

Case Study 2

Case Study



Adversarial Patch
https://arxiv.org/pdf/1712.09665.pdf

DataScience@SMU

Case Study 3

Case Study

The left image shows real graffiti on a Stop sign, something that most humans would not think is suspicious. The right image shows a physical perturbation applied to a Stop sign. We design our perturbations to mimic graffiti, and thus ”hide in the human psyche.”

Robust Physical-World Attacks on Deep Learning Visual Classification
https://arxiv.org/pdf/1707.08945.pdf

DataScience@SMU

Countering Adversarial Attacks

Countering Adversarial Examples

When the underlying optimization function is nonlinear and nonconvex (i.e., the function is difficult to solve directly by many ML models), the adversarial examples are solutions to such a function. Since we don't have theoretical tools to solve these types of functions, we cannot design an optimal adversarial defense either.

Adversarial attacks are difficult because they exploit a larger space of possible inputs than the algorithms are expecting to encounter.

Defenses Examples

- **Adversarial training:** generate multiple (types of) adversarial examples and train models explicitly so they are not affected by these adversarial examples.
- **Defensive distillation:** the model is trained to output probabilities for each class rather than hard boundaries for each class.

DataScience@SMU

Problematic Deep Learning

Problematic Deep Learning

- Deep learning allows the construction and identification of features such as race that may have previously been removed due to legal reasons.
- Deep learning very closely fits the data encoded in historical datasets. Historical datasets are often rife with encoded bias.
- Lack of minority involvement in training models can often miss culturally relevant examples that would be readily apparent to cultural stakeholders.
- Reverse engineering models to identify bias (an auditing approach) can be extremely hard, and oversight and insight is required at all levels of the process.

Future Reading: *Weapons of Math Destruction* by Cathy O'Neil

DataScience@SMU