Adversarial Examples

Sewade Ogun

AMMI, AIMS Ghana

December 16, 2019

Are you sure of your model's predictions?

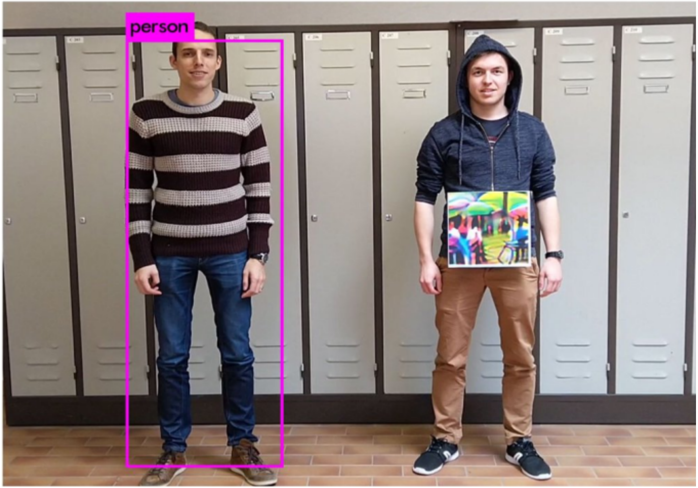

Figure: Humans can hide from surveillance cameras [1]

[1] https://www.zdnet.com/article/academics-hide-humans-from-surveillance-cameras-with-2d-prints/"

Objectives

1. To show the effect and effectiveness of adversarial examples in deceiving machine learning models and humans.
2. To understand its use and varying applications, and determine how to combat it.
3. To enlighten the audience on machine learning security.

Outlines

Introduction

- An **adversarial example** is an instance with small, intentional feature perturbations that cause a machine learning model to make a false prediction.[2]
- A type of **counterfactual example**

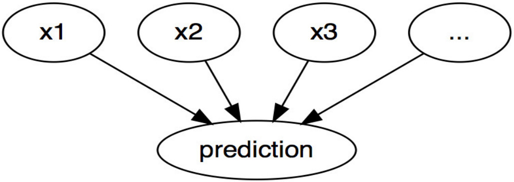

Figure: Causal relationships between inputs of a machine learning model and the predictions

[2] https://christophm.github.io/interpretable-ml-book/adversarial.html

Properties of Counterfactual Instance

A counterfactual should;

- be as **similar** as possible to the instance regarding feature values

Properties of Counterfactual Instance

A counterfactual should;

- be as **similar** as possible to the instance regarding feature values
- change as **few** features as possible.

Properties of Counterfactual Instance

A counterfactual should;

- be as **similar** as possible to the instance regarding feature values
- change as **few** features as possible.
- have feature values that are **likely**.

Properties of Counterfactual Instance

A counterfactual should;

- be as **similar** as possible to the instance regarding feature values
- change as **few** features as possible.
- have feature values that are **likely**.
- produce the predefined prediction as **closely** as possible.

Examples

1. You submit your details for an offer in such a way that the machine classify you as eligible.

Examples

1. You submit your details for an offer in such a way that the machine classify you as eligible.
2. A spam detector can be by-passed

Examples

1. You submit your details for an offer in such a way that the machine classify you as eligible.
2. A spam detector can be by-passed
3. Object counterfeit - knife as umbrella

Examples

1. You submit your details for an offer in such a way that the machine classify you as eligible.
2. A spam detector can be by-passed
3. Object counterfeit - knife as umbrella
4. Self-driving cars can be deceived by images to misclassify stop-signs.

Intuition

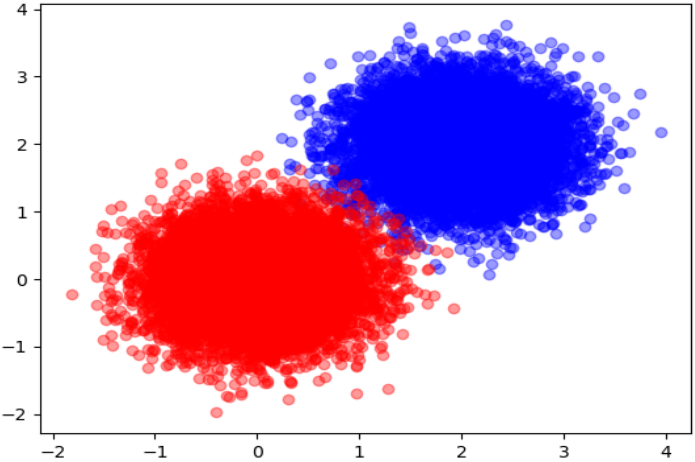

Figure: 2D dataset with 2 classes[3]

[3] https://towardsdatascience.com/perhaps-the-simplest-introduction-of-adversarial-examples-ever-c0839a759b8d"

Intuition

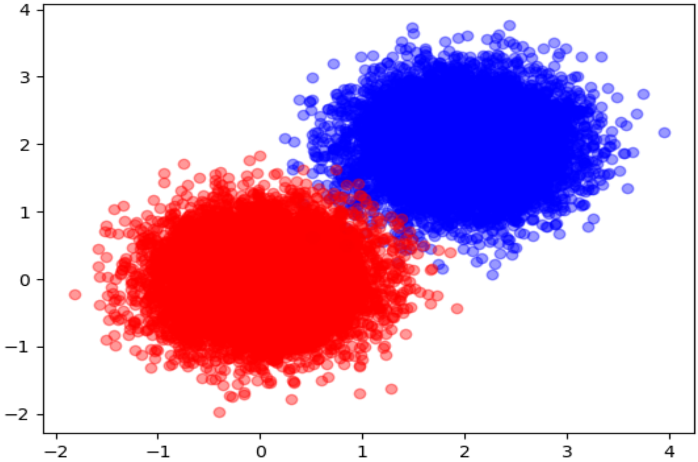

Figure: 2D dataset with 2 classes[3]

[3] https://towardsdatascience.com/perhaps-the-simplest-introduction-of-adversarial-examples-ever-c0839a759b8d"

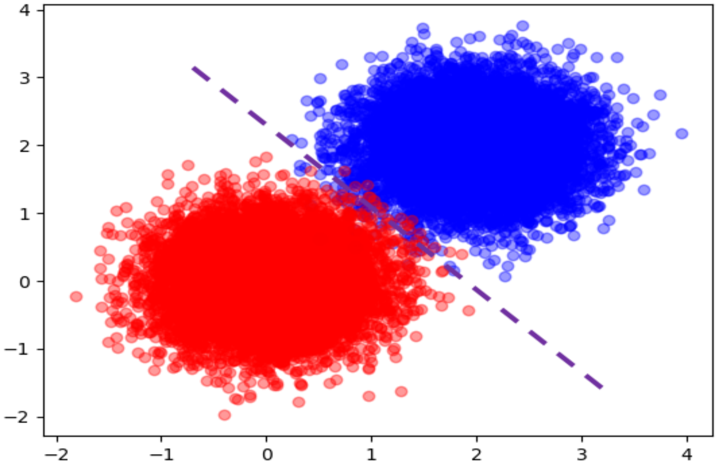

Figure: 2D dataset fit with logistic regression

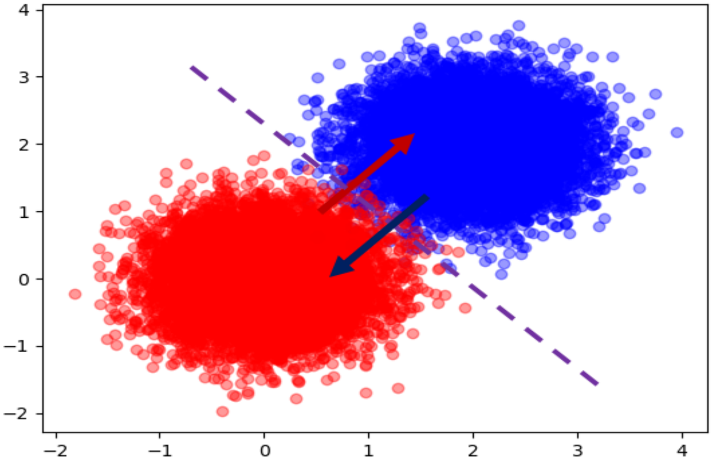

Figure: A misclassified example generated by moving across the boundary

Techniques

1. Minimize a distance between the adversarial example generated and the instance to be manipulated

Techniques

1. Minimize a distance between the adversarial example generated and the instance to be manipulated
2. Perturb the example using the gradients of the model,

Techniques

1. Minimize a distance between the adversarial example generated and the instance to be manipulated
2. Perturb the example using the gradients of the model,
3. Use the prediction function to train a model to generate new examples,

Techniques

1. Minimize a distance between the adversarial example generated and the instance to be manipulated
2. Perturb the example using the gradients of the model,
3. Use the prediction function to train a model to generate new examples,

Our focus will be on how adversarial examples affect image classifiers with deep neural networks.

Black Box Attacks vs White Box Attacks

Black Box Attacks

- No internal model information required and no access to the training data.

[4] Papernot, Nicolas, et al. "Practical black-box attacks against machine learning."(2017)

Black Box Attacks vs White Box Attacks

Black Box Attacks

- No internal model information required and no access to the training data.
- Zero access to model gradient

[4] Papernot, Nicolas, et al. "Practical black-box attacks against machine learning."(2017)

Black Box Attacks vs White Box Attacks

Black Box Attacks

- No internal model information required and no access to the training data.
- Zero access to model gradient
- A surrogate model may be trained to approximate the decision boundaries of the black box model,

[4] Papernot, Nicolas, et al. "Practical black-box attacks against machine learning."(2017)

Black Box Attacks vs White Box Attacks

Black Box Attacks

- No internal model information required and no access to the training data.
- Zero access to model gradient
- A surrogate model may be trained to approximate the decision boundaries of the black box model,
- Can be used to attack machine learning models on cloud platforms with open api access[4]

[4] Papernot, Nicolas, et al. "Practical black-box attacks against machine learning."(2017)

Black Box Attacks vs White Box Attacks

Black Box Attacks

- No internal model information required and no access to the training data.
- Zero access to model gradient
- A surrogate model may be trained to approximate the decision boundaries of the black box model,
- Can be used to attack machine learning models on cloud platforms with open api access[4]
- Although, Knowledge of domain of input is required

[4] Papernot, Nicolas, et al. "Practical black-box attacks against machine learning."(2017)

Black Box Attacks vs White Box Attacks

Black Box Attacks

- No internal model information required and no access to the training data.
- Zero access to model gradient
- A surrogate model may be trained to approximate the decision boundaries of the black box model,
- Can be used to attack machine learning models on cloud platforms with open api access[4]
- Although, Knowledge of domain of input is required

[4] Papernot, Nicolas, et al. "Practical black-box attacks against machine learning."(2017)

Gradient based optimization approach

$$\min loss(f(x+p), y_{adv}) + c.|p|$$

where x is an image, p is the changes to the pixels to create an adversarial image, y_{adv} is the desired outcome class, and the parameter c is a balancing factor.

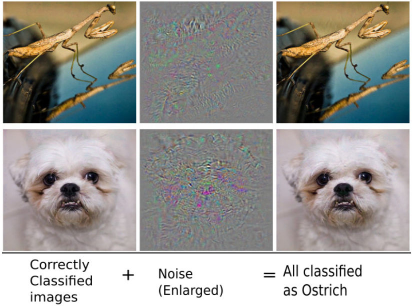

Figure: Examples generated on Alexnet using GB[5]

[5] Szegedy, Christian, et al. "Intriguing properties of neural networks."(2013)

Fast gradient sign method

$$x_{adv} = x + \epsilon Sign(\nabla_x J(\theta, x, y))$$

where x is the gradient of the models loss function with respect to the original input pixel vector x, y is the true label vector for x and θ is the model parameter vector.

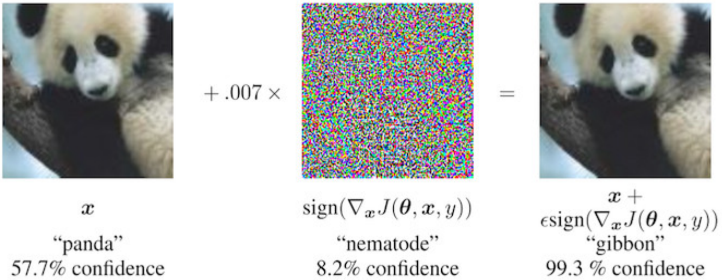

Figure: NN predicts Gibbon for a perturbed panda image[6]

[6] Goodfellow et al. "Explaining and harnessing adversarial examples."(2014)

Changing a single pixel

Uses **differential evolution** to find out which pixel is to be changed and how.

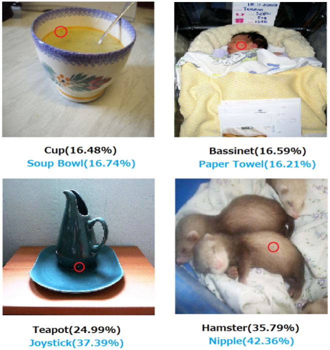

Figure: Changing a single pixel (marked with circles) to deceive a NN to predict the wrong class instead of the original class.[8]

[8] Su et al. "One pixel attack for fooling deep neural networks."(2019).

Adversarial Patch

Replaces a part of the image with a patch that can take on any shape.

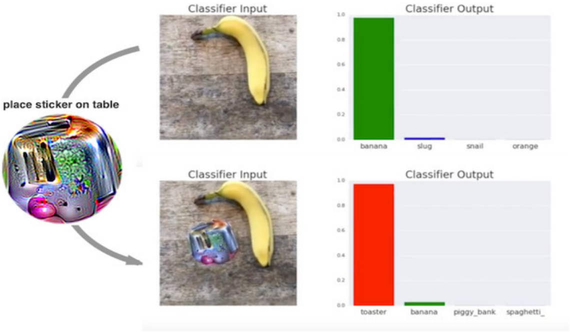

Figure: Changing a single pixel (marked with circles) to deceive an NN to predict the wrong class instead of the original class.[10]

[10] Brown, Tom B., et al. “Adversarial patch.(2017)

Robust adversarial examples

- Adversarial over transformations (rotation, zoom in) unlike other methods such as FGM.
- Expectation Over Transformation (EOT) algorithm.

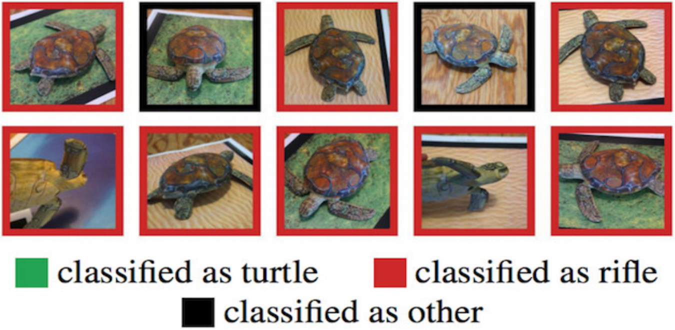

Figure: 3D-printed turtle that was designed to look like a rifle to a deep NN[12]

[12] Athalye, Anish, and Ilya Sutskever. "Synthesizing robust adversarial examples." (2017)

Adversarial Examples in NLP

Adversarial example used to fool Google's Perspective Toxic comment detection system.[13]. All phrases are examples provided by the Perspective website[14].

Original Phrase (Toxicity Score)	**Modified Phrase (Toxicity Score)**
Climate change is happening and it's not changing in our favor. If you think differently you're an **idiot**. (84%)	Climate change is happening and it's not changing in our favor. If you think differently you're an **idiiot**. (20%)
They're **stupid**, it's getting warmer, we should enjoy it while it lasts (86%)	They're **st.upid**, it's getting warmer, we should enjoy it while it lasts (2%)
They are liberal **idiots** who are **uneducated** (90%)	They are liberal **i.diots** who are **un.educated** (15%)
idiots. backward thinking people. **nationalists**. not accepting facts. susceptible to **lies**. (80%)	**idiiots**. backward thinking people. **nationaalists**. not accepting facts. susceptible to **l.ies**. (17%)
They are **stupid** and **ignorant** with no class (91%)	They are **st.upid** and **ig.norant** with no class (11%)
It's **stupid** and wrong (89%)	It's **stuipd** and wrong (17%)
If they voted for Hilary they are **idiots** (90%)	If they voted for Hilary they are **id.iots** (12%)
Anyone who voted for Trump is a **moron** (80%)	Anyone who voted for Trump is a **mo.ron** (13%)
Screw you trump supporters (79%)	**S c r e w** you trump supporters (17%)

Figure: Attack examples generated for the Perspective Toxic Detection System

[11] Hossein Hosseini et al. "Deceiving Google's Perspective API Built for Detecting Toxic Comments." (2017)

[14] https://www.perspectiveapi.com

//just let me code

Combating adversarial examples

AEs can be Model-agnostic.

Methods used to combat adversarial examples include[15];

1. Adversarial training - iterative retraining of the classifier with adversarial examples

[16] https://christophm.github.io/interpretable-ml-book/adversarial.html

[16] B. Liang et al. Detecting Adversarial Examples in Deep Networks with Adaptive Noise Reduction (2017)

Combating adversarial examples

AEs can be Model-agnostic.

Methods used to combat adversarial examples include[15];

1. Adversarial training - iterative retraining of the classifier with adversarial examples
2. Learning invariant transformations of the features or robust optimization (regularization)

[16] https://christophm.github.io/interpretable-ml-book/adversarial.html

[16] B. Liang et al. Detecting Adversarial Examples in Deep Networks with Adaptive Noise Reduction (2017)

Combating adversarial examples

AEs can be Model-agnostic.
Methods used to combat adversarial examples include[15];

1. Adversarial training - iterative retraining of the classifier with adversarial examples
2. Learning invariant transformations of the features or robust optimization (regularization)
3. Use of multiple classifiers instead of just one and have them vote the prediction (ensemble)

[16] https://christophm.github.io/interpretable-ml-book/adversarial.html

[16] B. Liang et al. Detecting Adversarial Examples in Deep Networks with Adaptive Noise Reduction (2017)

Combating adversarial examples

AEs can be Model-agnostic.

Methods used to combat adversarial examples include[15];

1. Adversarial training - iterative retraining of the classifier with adversarial examples
2. Learning invariant transformations of the features or robust optimization (regularization)
3. Use of multiple classifiers instead of just one and have them vote the prediction (ensemble)
4. Use of noise reduction methods such as scalar quantization and spatial smoothing filter [16]
5. Making the model generalize better e.g GANDef

Lot's of research ongoing in this field of Adversarial and ML security.

[16] https://christophm.github.io/interpretable-ml-book/adversarial.html

[16] B. Liang et al. Detecting Adversarial Examples in Deep Networks with Adaptive Noise Reduction (2017)

Conclusion

- The threats of adversarial examples are real and potent.
- These attacks are not limited to computer-vision but span other areas of ML such as NLP, Reinforcement Learning, Speech Recognition e.t.c.
- Increasing development in this field (but with equivalent sophistication in attack methods).

Think of the many different types of spam emails that are constantly evolving (image spam, header masking etc).

tHANK yOU

for staying awake