UNIVERSIDADE DA CORUÑA

INESCTEC

Are Neural Networks secure? Introduction to Adversarial Attacks

Brais Cancela[1,2]

June 13th, 2019

[1]LIDIA Group, Universidade da Coruña
[2]LIAAD Group, University of Porto

DSPT

DATA SCIENCE PORTUGAL

Content

Introduction

Introduction

Mark the differences

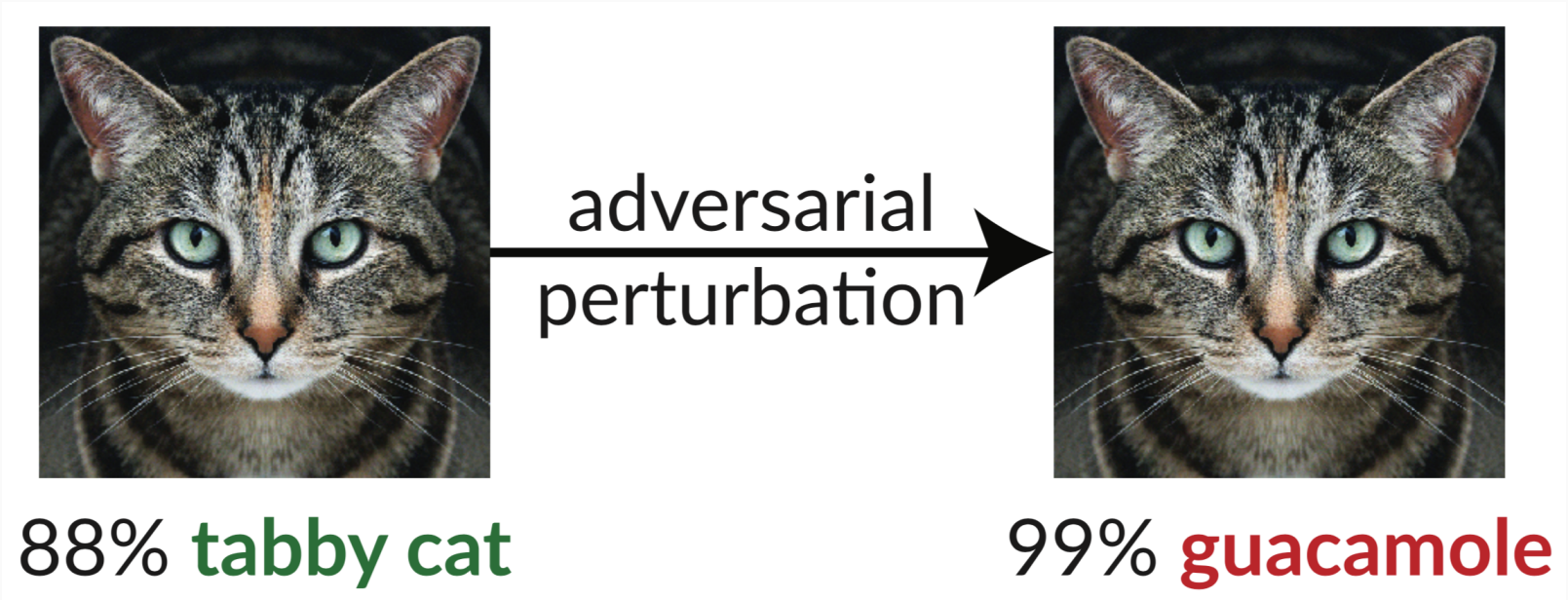
Introduction

Mark the differences

88% tabby cat

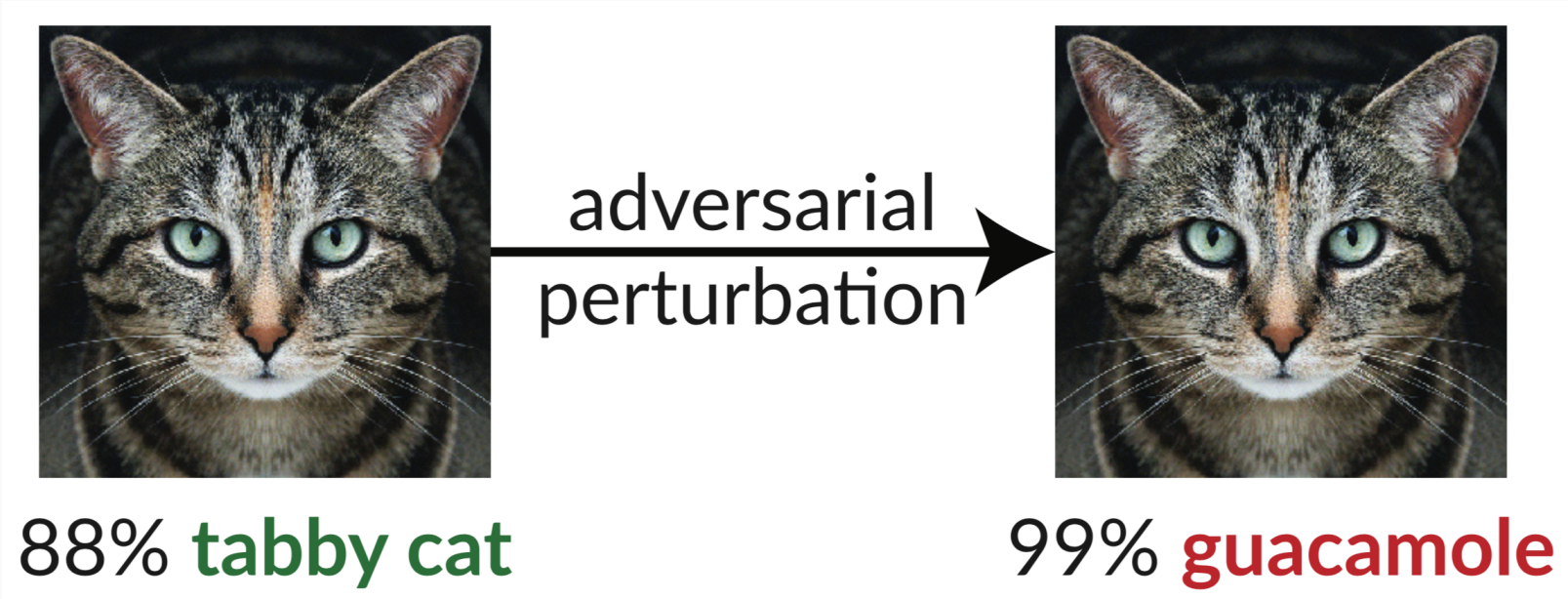
99% guacamole

Mark the differences



Introduction

Mark the differences

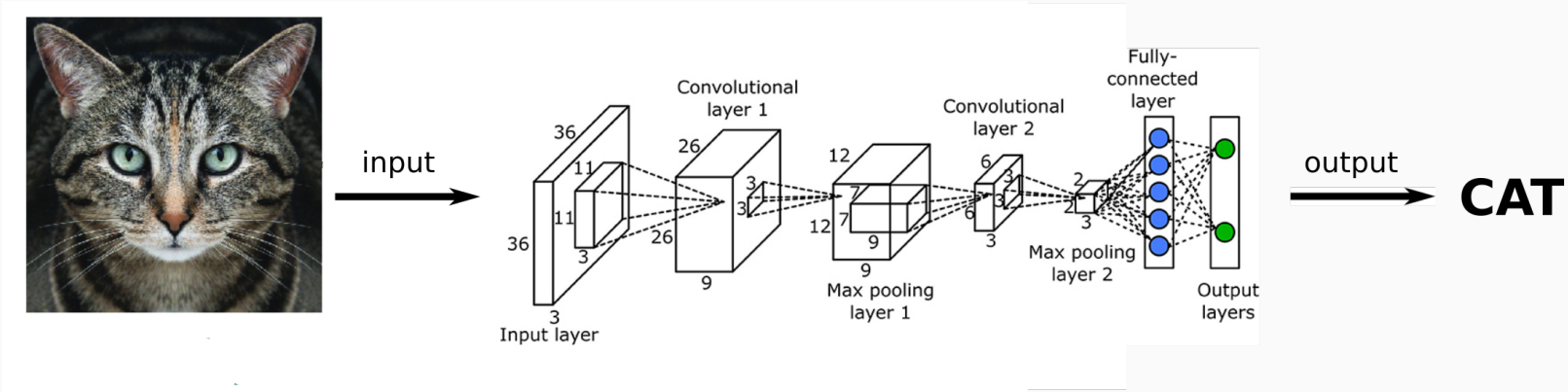


How easy is too fool a CNN?

Adversarial Attack

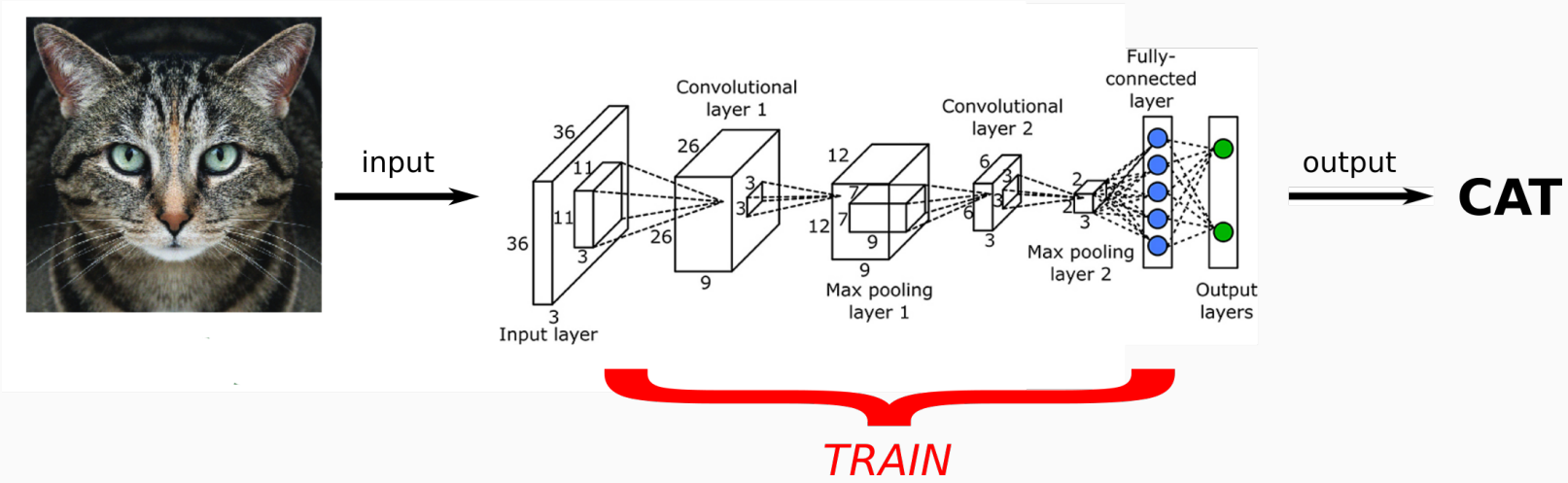
Adversarial Attack

White-box Attack



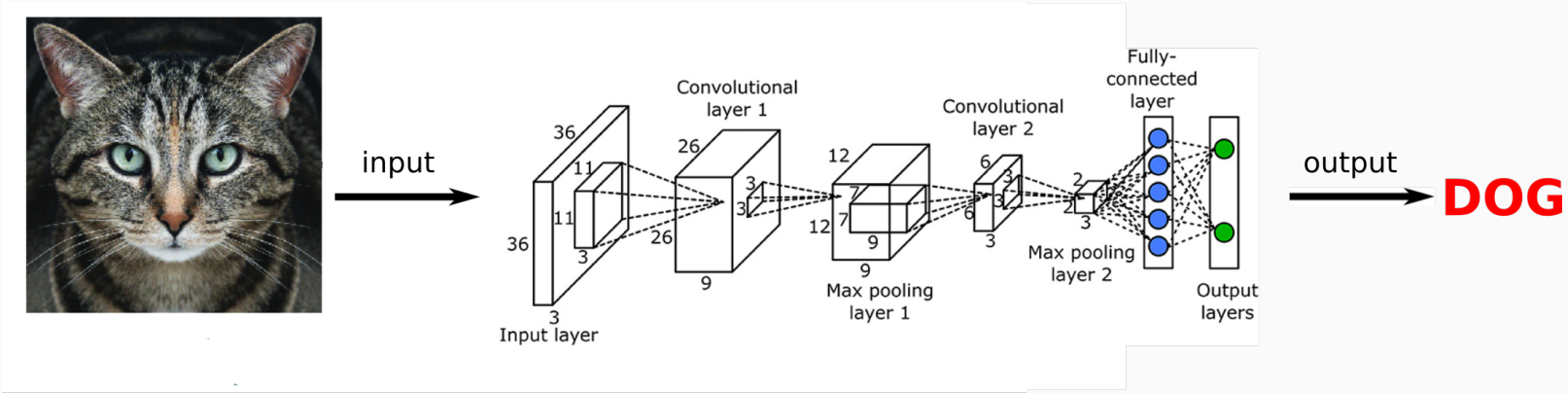
Adversarial Attack

White-box Attack



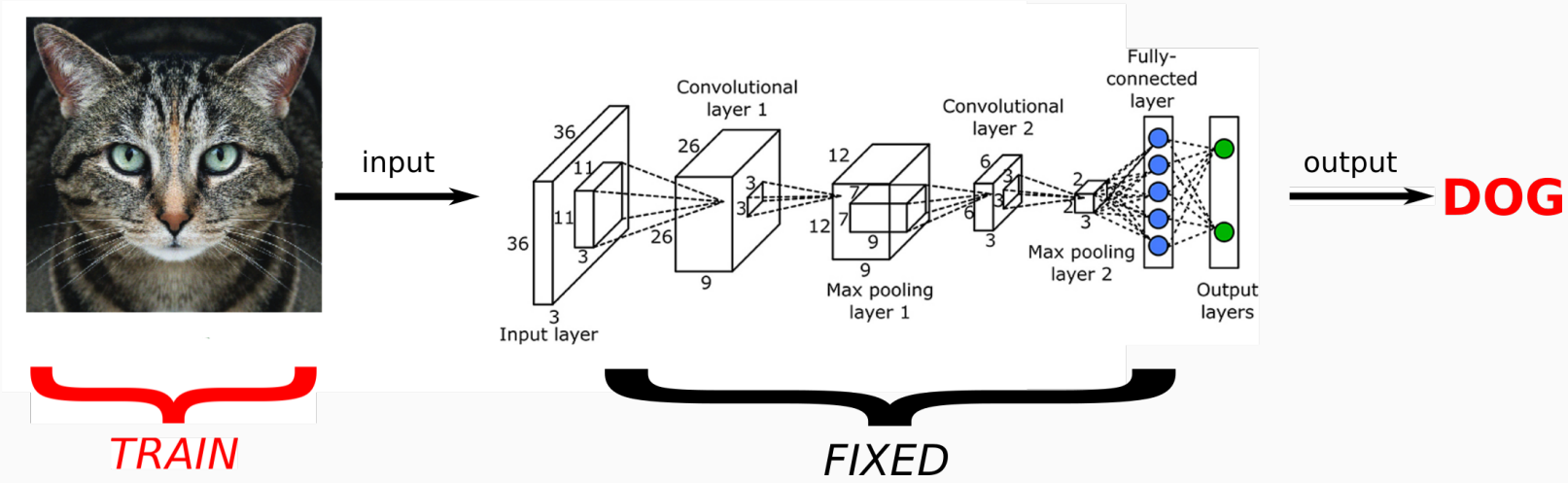
Adversarial Attack

White-box Attack



Adversarial Attack

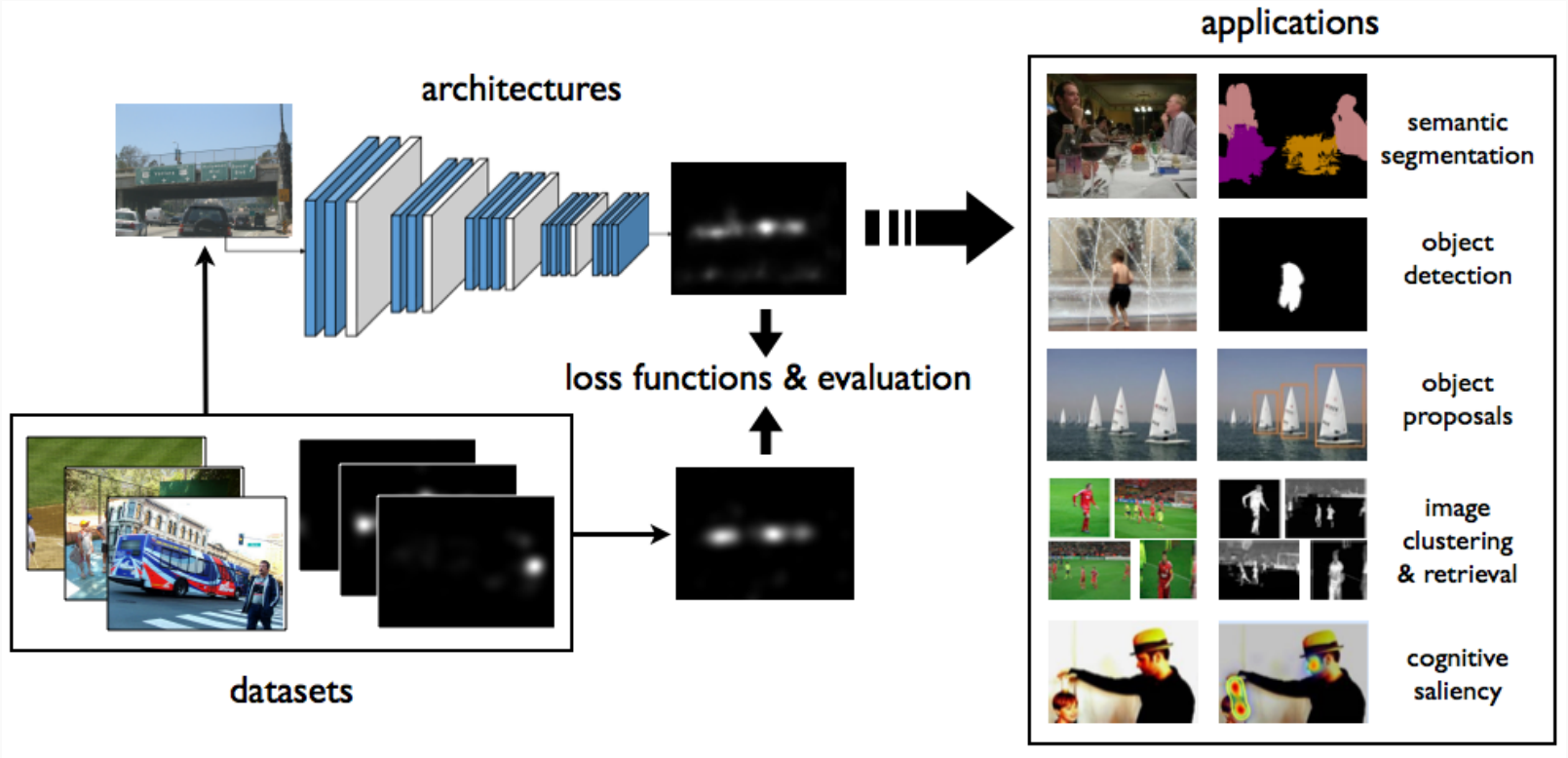
White-box Attack



Saliency

Appears for the first time in Computer Vision problems [1]

Pointed out how features are contributing to the classifier's output.



[1] K. Simonyan, A. Vedaldi y A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps", 2013

Adversarial Attack

Add *guided* noise to the input

Use the image gradient to establish the direction.

Adversarial Methods

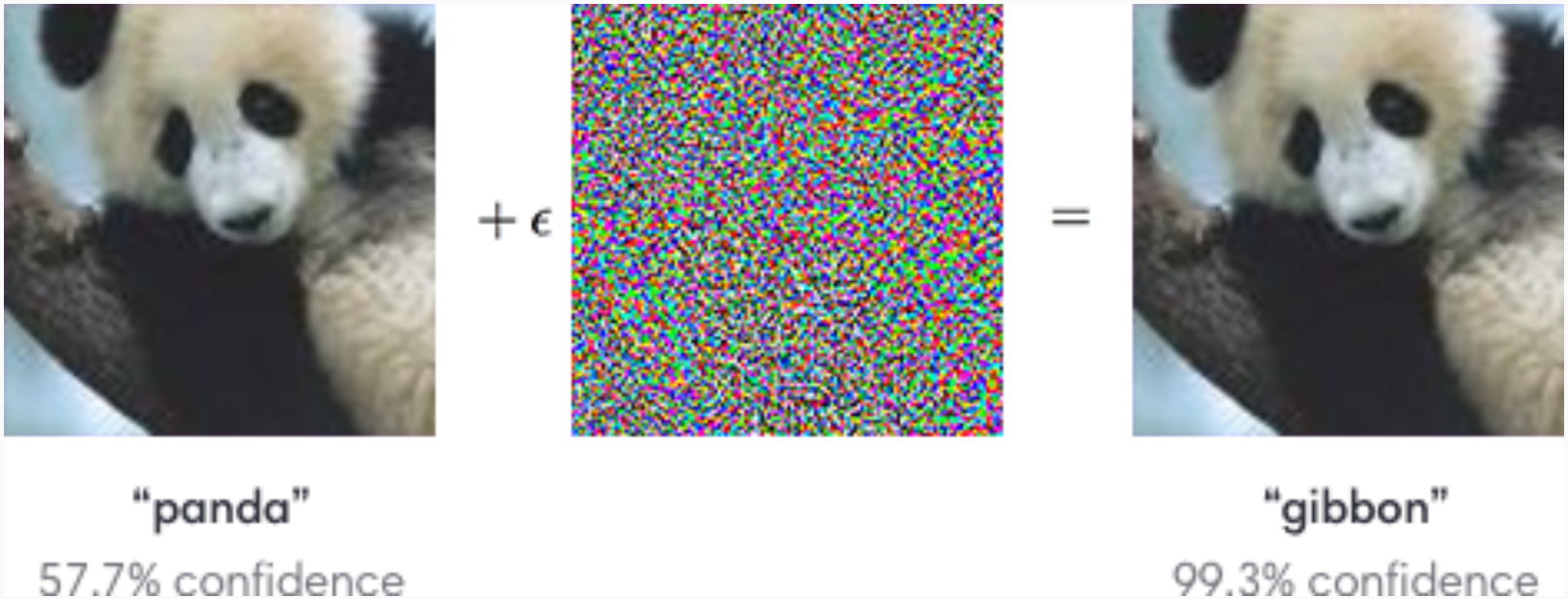
- Fast Gradient Sign Method (FGSM)
- FGSM variants (I-FGSM, FGSM-LL, . . .)
- Projected Gradient Descent (PGD)

Adversarial Attack

Add *guided* noise to the input

Use the image gradient to establish the direction.

How it works



Adversarial Attacks

DoS Attack

- The real class must never win
- In security environments, the idea is to force an individual to never log in into the system.

Phishing Attack

- A predefined class must win with a high confidence (high probability)
- In security environments, the idea is to force an unknown individual to log into the system with a specific user's credentials.

How to prevent it

How to prevent it

Hide the Model

Hide the model architecture, so it is not known by the attacker.

How to prevent it

Hide the Model

Hide the model architecture, so it is not known by the attacker.

Is it enough? Transfer learning

How to prevent it

Hide the Model

Hide the model architecture, so it is not known by the attacker.

Hide the Data

Hide the data used to train the model.

How to prevent it

Hide the Model

Hide the model architecture, so it is not known by the attacker.

Hide the Data

Hide the data used to train the model.

Is it enough? Adversarial attack

How to prevent it

Hide the Model

Hide the model architecture, so it is not known by the attacker.

Hide the Data

Hide the data used to train the model.

Adversarial Training

Create images that fool the classifier, then add it to the training procedure.

How to prevent it

Hide the Model

Hide the model architecture, so it is not known by the attacker.

Hide the Data

Hide the data used to train the model.

Adversarial Training

Create images that fool the classifier, then add it to the training procedure.

Is it enough? Accuracy drop

How to prevent it

Hide the Model

Hide the model architecture, so it is not known by the attacker.

Hide the Data

Hide the data used to train the model.

Adversarial Training

Create images that fool the classifier, then add it to the training procedure.

Add Noise

Add Gaussian noise to the image input.

How to prevent it

Hide the Model

Hide the model architecture, so it is not known by the attacker.

Hide the Data

Hide the data used to train the model.

Adversarial Training

Create images that fool the classifier, then add it to the training procedure.

Add Noise

Add Gaussian noise to the image input.

Is it enough? Smooths the error surface

How to prevent it

Hide the Model

Hide the model architecture, so it is not known by the attacker.

Hide the Data

Hide the data used to train the model.

Adversarial Training

Create images that fool the classifier, then add it to the training procedure.

Add Noise

Add Gaussian noise to the image input.

Use a discriminator

A classifier confirms whether the input is a real image or a fake one.

How to prevent it

Hide the Model

Hide the model architecture, so it is not known by the attacker.

Hide the Data

Hide the data used to train the model.

Adversarial Training

Create images that fool the classifier, then add it to the training procedure.

Add Noise

Add Gaussian noise to the image input.

Use a discriminator

A classifier confirms whether the input is a real image or a fake one.

Is it enough? Adversarial attack over the discriminator

Conclusion

Conclusion

To sum up

- Adversarial attacks are a big concern.
- It is a white-box attack, but can work even when the model architecture is hidden.
- Hiding architecture + input noise are the best strategy right now.

Future Directions

- Novel architectures.
- Novel training functions.

Thanks!