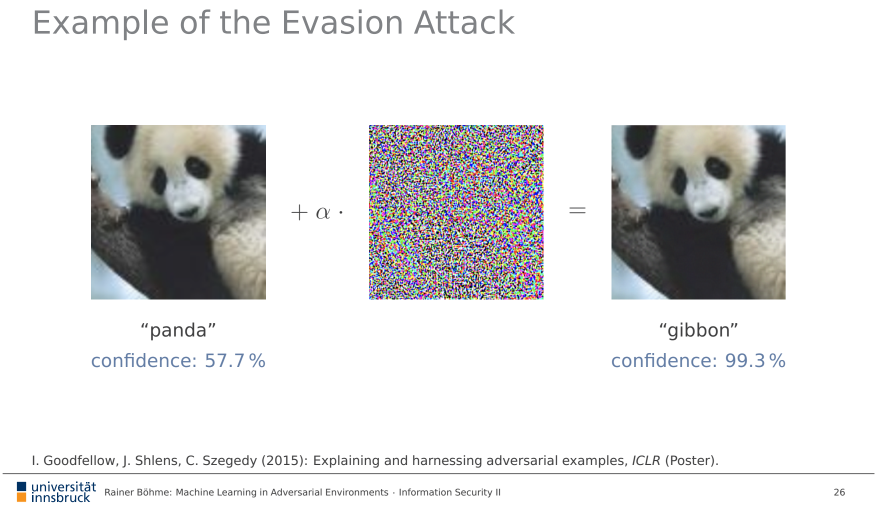
universität
innsbruck

Adversarial Label Flips

Matthias Dellago & Maximilian Samsinger

Previously on InfoSec 2...



Example of the Evasion Attack

$+ \alpha \cdot$ $=$

“panda”

confidence: 57.7 %

“gibbon”

confidence: 99.3 %

I. Goodfellow, J. Shlens, C. Szegedy (2015): Explaining and harnessing adversarial examples, *ICLR* (Poster).

universität innsbruck Rainer Böhme: Machine Learning in Adversarial Environments · Information Security II 26

Idea

Evasion Attack

Use backpropagation with two significant differences:

1. change input values, instead of weights and biases
2. increase cost function, instead of decrease

DeepDream

DeepDream applies 1, but not 2. So we are doing modified DeepDreaming in a sense.

Expected Outcome

Confusion Matrix

		Categorised as		
		Dog	Cat	Plane
Adversarial Example of a	Dog	0.0	?	?
	Cat	?	0.0	?
	Plane	?	?	0.0

Hypothesis

Uniform Distribution?

- Is post-attack label uniformly distributed over all other labels (null hypothesis) or not?
- If not, why?

Methods

Datasets

MNIST, Fashion MNIST, CIFAR-10

Models

ResNet-18 for CIFAR-10. Some simple convolutional neural network for MNIST & Fashion MNIST.

Attacks

FGSM and PGD

Stretch goals

- Reverse Deep Dreaming
- Consider natural adversarial examples
- Think about applications (attacker and defender)
- More attacks and/or architectures

Brainstorming slide (will be removed)

What do you want to achieve till the end of semester?

1. Investigate relationship between ground truth labels and predicted label of adversarial examples. (Maybe formulate as null hyposisis: No correlation)
2. Github repo for reproducibility.
3. Max. Learn PyTorch
4. Matthias. Learn ML

Why is your topic relevant?

Contribution to basic research.