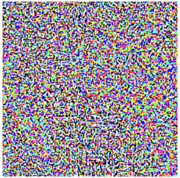
Adversarial Label Flips

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).

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, in a sense, we are doing modified DeepDreaming.

Expected Outcome

Confusion Matrix

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

How many modified dogs get classified as cats vs as planes? etc.

Hypothesis

Uniform Distribution?

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

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 DeepDreaming: What does exaggerated evasion look like?
- Think about applications, use of confusion matrix (attacker and defender)
- More attacks and/or architectures
- Natural adversarial examples
- Adversarially robust networks (L. Schott et al.)