universität
innsbruck

Adversarial Label Flips

Matthias Dellago & Maximilian Samsinger

Introduction

Evasion Attack

As seen in Lecture. Insert lecture slides or explain anew?

Introduction

Confusion Matrix

Adverserial Example		Categorised as		
		Dog	Cat	Plane
	Dog	0.0	?	?
	Cat	?	0.0	?
	Plane	?	?	0.0

Remove

Summarize what #5 is about and recap what the students need to know. (Maybe needs an extra slide) Explain (or visualize?) confusion matrix.

Hypothesis

Uniform Distribution?

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

Reasons?

- If uniform, why? If not, why not? Possible relationships between different classes.
- Probably intractable, but interesting.

Methods

Datasets

MNIST, Fashion MNIST, CIFAR-10

Alert block

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

Attacks

FGSM and PGD

Stretch goals

- Study natural adversarial examples
- Look for 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.