universität
innsbruck

Adversarial Label Flips

Matthias Dellago & Maximilian Samsinger

Introduction

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

Hypothesis

Research question

Does a relationship between ground truth and final predicted label exist given an untargeted attack?

First question

Does a relationship exist *at all*.

Null hypothesis

The final predicted label is randomly (uniform) drawn from the set of all other labels.

Alternative hypothesis

The final predicted label is not randomly (uniform) drawn from the set of all other labels.

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

Kinda ordered by relevancy

1. Study natural adversarial examples
2. Looking for applications
3. 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.

Stretch goals

More models/architectures.
Influence of adversarial training.

Example Slide

Null hypothesis

Adversarial examples

Alert block

This is a alertblock

Example block

Examples show up one by one. . . .

- Example 1

Example Slide

Null hypothesis

Adversarial examples

Alert block

This is a alertblock

Example block

Examples show up one by one. . . .

- Example 1
- Example 2

Example Slide

Null hypothesis

Adversarial examples

Alert block

This is a alertblock

Example block

Examples show up one by one. . . .

- Example 1
- Example 2
- Example 3