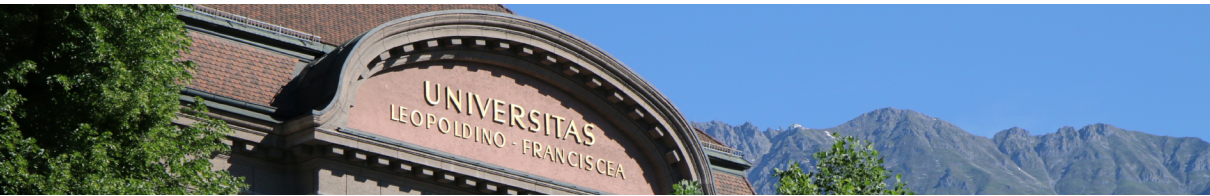
universität
innsbruck



Adversarial Label Flips

Matthias Dellago & Maximilian Samsinger

Introduction/Reminder from last time

Panda slide from last time

Introduction/Reminder from last time

What do we want to do?

Generate adversarial examples.
Create confusion matrix

Confusion matrix

Confusion matrix from last presentation

Adversarial examples + some attacks

Block

Adversarial examples have been introduced in [1].

Examples of attacks: FGSM + PGD

FGSM has been introduced in [2]. (Panda example)
An iterated version (PGD) was introduced in [3].

How does FGSM work?

FGSM

text

How does FGSM work?

FGSM

text

For more attacks we use Foolbox.

What is Foolbox?

Foolbox

A suit of attacks is available with FoolBox! [4]. Maybe show the full list of attacks directly on the website. `https://foolbox.readthedocs.io/en/stable/modules/attacks.html`

Implementation

Foolbox can be used with PyTorch, TensorFlow and JAX. We arbitrarily choose Pytorch!

Optional Slide 1: Data set

Some images for MNIST, Fashion-MNIST and CIFAR-10.

Optional Slide 2: Convolutional neural networks

We use small convolutional neural networks [5] for the "easy" data sets. For CIFAR-10 we will use ResNet-18, a residual neural network [6], [7].

References I

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus.
Intriguing properties of neural networks.
In *International Conference on Learning Representations (ICLR)*, 2014.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy.
Explaining and harnessing adversarial examples.
arXiv preprint arXiv:1412.6572, 2014.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu.
Towards deep learning models resistant to adversarial attacks.
arXiv preprint arXiv:1706.06083, 2017.

References II

- Jonas Rauber, Wieland Brendel, and Matthias Bethge. Foolbox: A python toolbox to benchmark the robustness of machine learning models. *arXiv preprint arXiv:1707.04131*, 2017.
- Yann LeCun, Patrick Haffner, Léon Bottou, and Yoshua Bengio. Object recognition with gradient-based learning. In *Shape, contour and grouping in computer vision*, pages 319–345. Springer, 1999.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

References III

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun.
Identity mappings in deep residual networks.
In *European Conference on Computer Vision*, pages 630–645.
Springer, 2016.