universität
innsbruck


# Adversarial Label Flips

Related work presentation

Matthias Dellago & Maximilian Samsinger

# Standard sources on adversarial examples

**Adversarial examples [1]**

Deep neural networks are vulnerable to adversarial perturbations!

[1] Intriguing properties of neural networks, 2014
[2] Explaining and harnessing adversarial examples, 2014

# Standard sources on adversarial examples

**Adversarial examples [1]**

Deep neural networks are vulnerable to adversarial perturbations!

**Fast gradient sign method [2]**

Even worse: Adversarial examples generalize over multiple datasets and architectures. Even cheap attacks like FGSM can work.

[1] Intriguing properties of neural networks, 2014
[2] Explaining and harnessing adversarial examples, 2014

# Standard sources on adversarial examples

**Adversarial examples [1]**

Deep neural networks are vulnerable to adversarial perturbations!

**Fast gradient sign method [2]**

Even worse: Adversarial examples generalize over multiple datasets and architectures. Even cheap attacks like FGSM can work.

**Fast gradient sign method**

Modify an input image $x$, with respective label $y$,

$$x + \epsilon \, \mathrm{sign}(\nabla_x J(\theta, x, y)).$$

using the loss function $J$.

[1] Intriguing properties of neural networks, 2014
[2] Explaining and harnessing adversarial examples, 2014

# Standard sources on adversarial examples

**Adversarial examples [1]**

Deep neural networks are vulnerable to adversarial perturbations!

**Projected gradient descent [3]**

Strong and cheap attacks therefore exist: Projected gradient descent (PGD) iteratively computes FGSM.

**Fast gradient sign method**

Modify an input image $x$, with respective label $y$,

$$x + \epsilon \, \mathrm{sign}(\nabla_x J(\theta, x, y)).$$

using the loss function $J$.

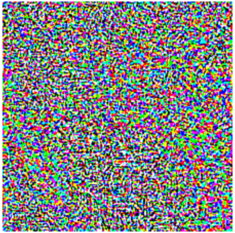
[1] Intriguing properties of neural networks, 2014
[3] Towards deep learning models resistant to adversarial attacks, 2018

# Fast gradient sign method

Panda

(57.7% confidence)

$+ \epsilon$



$\text{sign}(\nabla_x J(\theta, x, y))$

$=$

Gibbon

(99.3% confidence)

[2] Explaining and harnessing adversarial examples, 2014

# What we want to do

**Confusion Matrix**

Categorised as

| Adversarial Example of a | | Dog | Cat | Plane |
|---|---|---|---|---|
| | Dog | 0.0 | ? | ? |
| | Cat | ? | 0.0 | ? |
| | Plane | ? | ? | 0.0 |

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

# Case study

# Foolbox

A suit of attacks is available with FoolBox! [4]

[4] Foolbox: A python toolbox to benchmark the robustness of machine learning models, 2017

# Foolbox

A suit of attacks is available with FoolBox! [4]

**Foolbox**

- Over 40 different attacks.
- Available in PyTorch, TensorFlow and JAX.
- Easy to work with.

[4] Foolbox: A python toolbox to benchmark the robustness of machine learning models, 2017

# Foolbox

Projected Gradient Descent (PGD) attack for different `epsilons`.

[4] Foolbox: A python toolbox to benchmark the robustness of machine learning models, 2017

# Foolbox

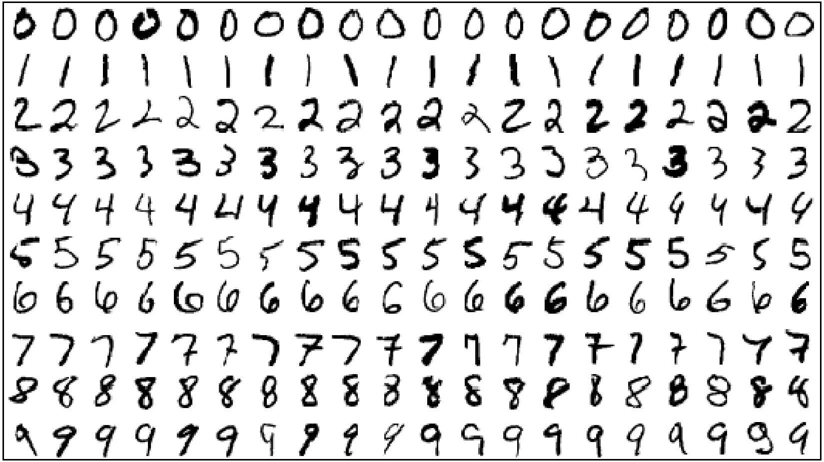
Projected Gradient Descent (PGD) attack for different `epsilons`.

```
import foolbox as fb

model = ...
fmodel = fb.PyTorchModel(model, bounds=(0, 1))

attack = fb.attacks.LinfPGD()
epsilons = [0.0, 0.001, 0.01, 0.03, 0.1, 0.3, 0.5, 1.0]
_, advs, success = attack(fmodel, images, labels, epsilons=epsilons)
```

[4] Foolbox: A python toolbox to benchmark the robustness of machine learning models, 2017

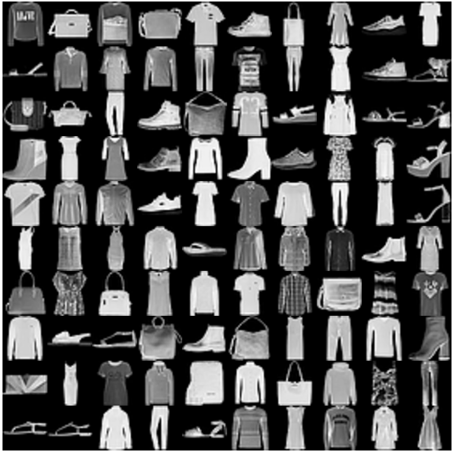
# Datasets

# MNIST [5]



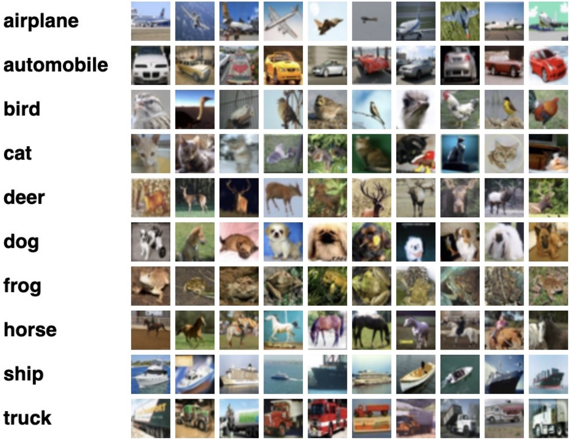
The MNIST database of handwritten digit images for machine learning research, 2012

# Fashion-MNIST [6]



Fashion-MNIST: A novel image dataset for benchmarking machine learning algorithms, 2017

# CIFAR-10 [7]




Learning multiple layers of features from tiny images, 2009

# 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.

Li Deng.
The mnist database of handwritten digit images for machine learning research [best of the web].
*IEEE Signal Processing Magazine*, 29(6):141–142, 2012.

Han Xiao, Kashif Rasul, and Roland Vollgraf.
Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms.
*arXiv preprint arXiv:1708.07747*, 2017.

# References III


Alex Krizhevsky, Geoffrey Hinton, et al.
Learning multiple layers of features from tiny images.
2009.