universität innsbruck

Adversarial Label Flips

Matthias Dellago & Maximilian Samsinger

Standard sources on adversarial examples

Adversarial examples

Adversarial examples have been introduced in [1].

Fast gradient sign method

FGSM has been introduced in [2].

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

Standard sources on adversarial examples

Adversarial examples

Adversarial examples have been introduced in [1].

Fast gradient sign method

FGSM has been introduced in [2].

Fast gradient sign method

Modify an input image x

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

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

Adversarial examples have been introduced in [1].

Fast gradient sign method

FGSM has been introduced in [2].

Fast gradient sign method

Modify an input image x

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

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

Adversarial examples have been introduced in [1].

Projected gradient descent

Projected gradient descent is a popular, strong attack, which iteratively computes FGSM. It was introduces in [3].

Fast gradient sign method

Modify an input image x

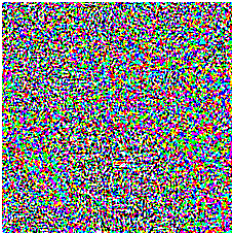
$$x + \epsilon \operatorname{sign}(\nabla_x J(\theta, x, y)).$$

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

$+ \quad \epsilon$



$=$

Panda
(57.7% confidence)

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

Gibbon
(99.3% confidence)

What we want to do

Confusion Matrix

Adversarial Example of a

		Categorised as	
	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

What is Foolbox?

Foolbox

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

Website

`https://foolbox.readthedocs.io`

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

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.