



Related work

# Adversarial Label Flips

Matthias Dellago & Maximilian Samsinger

April 21, 2021

# 1 On adversarial attacks

## 1.1 Attacks

**Existence of adversarial examples** Demonstrated that attacking deep neural networks are susceptible to attacks [1]. They actually coined the term "adversarial examples".

**Fast gradient sign method** [2] developed the fast gradient sign method. They are the guys with the panda image. Modify an input image  $x$ , with respective label  $y$ ,

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

using the loss function  $J$ .

**Projected gradient descent** The projected gradient descent, which is basically iterated FGSM, was first shown in [3]. Their experiments suggest that these attacks converge, i.e. they find a local maxima. This may require some restarts.

**Foolbox** A Python library with lots of attacks [4]. They include the attacks above.

# 2 On neural networks

## 2.1 Neural networks

First introduced in [5]. The authors of [6] demonstrated the effectiveness of deep convolutional neural networks on ImageNet.

**ResNets** Paradigm shift in deep learning. In [7] they developed Residual Networks to train very deep neural networks. We will probably use ResNet18. If we do, we probably also cite [8] for the "pre-activation" optimization. This is just a better architecture obtained by having BatchNorm-ReLU-Weights blocks instead of Weights-BatchNorm-ReLU blocks.

# 3 Methods

## 3.1 Datasets

MNIST, Fashion MNIST, CIFAR-10

## 3.2

We probably use <https://arxiv.org/pdf/1608.04644.pdf> Table 1 as a neural network for MNIST & Fashion-MNIST.

## References

- [1] 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.
- [2] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.
- [3] 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.
- [4] 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.
- [5] 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.
- [6] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 1097–1105, 2012.
- [7] 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.
- [8] 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.