Seminar thesis

# Adversarial Label Flips (sexier title?)

**Matthias Dellago & Maximilian Samsinger**

June 14, 2021

**Abstract**

Given a neural network (NN) trained to classify and a untargeted evasion attack [], in what class does the adversarial example fall? In the following, we will answer this question by evaluating some state of the art attacks, on a simple NN trained on industry standard datasets [**?**, **?**, **?**, **?**]. We discover a intuitive symmetry in the classification of adversarial examples.

# 1 Introduction

Adversarial examples exist. Targeted and untargeted. What does that look like? We use confusion matrices. example. how did we do it? -¿ NN + foolbox + mnist cifar (section methods) what did we get out of it? -¿ symmetric matrices and interesting insight about noise. (section results)

In 2013 Szegedy et al. demonstrated that attacking deep neural networks (NN) are susceptible to attacks [2]. These adversarial examples consist of a small perturbation applied to a otherwise innocuous input, engineered to cause the NN to misbehave.

In our case, the inputs will be images and the attack will be small changes to said image, designed to cause the classifier NN to misclassify the target.

These attacks on classifiers come in two different variations: targeted and untargeted. In targeted attacks[] the attacker aims to have the adversarial example identified as a specific class by the NN. (Say, misclassify a dog as a cat.) An untargeted attack meanwhile, only tries to evade correct classification. (Make this dog appear as anything, apart from a dog.)

Now, when considering untargeted attacks, the question as what the adversarial image *actually* is then classified, naturally arises. This is what we will experimentally answer in this paper.

We will present our results in terms of confusion matrices. In figure !! you can see an example. In larger matrices numbers become more difficult to process, so we will display our results in heatmap-style images (fig. !!). (Section results)

In our experiments we used the foolbox framework [], and simple NNs trained on the MNIST, FashionMNIST, and CIFAR-10 datasets []. We applied three state of the art attacks: Projected Gradient Decent (PDG)[], Carlini-Wagner [] and ...-Bethge[].

We show that, for the CIFAR-10 dataset the confusion matrices are surprisingly symmetric, and intuitively similar classes are often confused with each other. Furthermore we observe that for attacks which can choose large pertubations, there exist certain attractor classes, which most of the adversarial images are classified as. (Section Results.)

# 2 Background and related work

**Existence of adversarial examples** Since being first demonstrated [2] a large body of literature has flourished around adversarial examples. Blablabla...

**Notions of similarity** As mentioned previously we want to fool the NN with a small perturbation, that is to say a image which is very similar. But what do we precisely mean by "small pertubation" and similar? Here, we will briefly explain the most common methods to quantify this.

The distance of two images ($x$,$y$) (i.e. vectors) is described by a metric $D$. This metric can be defined via different norms, most commonly the $L^\infty$-, $L^1$- and $L^2$-norms.( The $L^0$-"norm" is also frequently used, though not a norm in the strict sense.) The distance between these the two images, is then defined as the $L^p$-norm of their difference, for a given $p$:

$$D(x, y) := \|y - x\|_p \tag{1}$$

### 2.1 Attacks

**Fast gradient sign method** [3] developed the fast gradient sign method. They are the guys with the panda image.

**Projected gradient descent** The projected gradient descent, which is basically iterated FGSM, was first shown in [1]. Their experiments suggest that these attacks converge, i.e. they find a local maxima. This may require some restarts.

**Carlini-Wagner attack**

**Bethge attack**

## 3 Methods

Reference to our github.

### 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.

## 4 Experiments

## 5 Results

pictures, pictures, and maybe a graph or two

## 6 Discussion

Symmetry of matrices -¿ maybe find a way to quantify symmetry?
-¿ NN can recognise "similarity"

Attractor classes -¿ manage with an extra "noise"-class or so?

In Figures X, Y and Z one can observe that adversarial examples computed with large perturbation budgets $\epsilon$ are misclassified as "8", "TODO" and "frog" for MNIST, Fashion-MNIST and CIFAR-10 respectively. In order to shed light onto this phenomenon we generate and classify 10000 white noise images sampled from a uniform distribution on the input domain. Figure A shows that these randomly generated images are also, most commonly, classified as "8", "TODO" and "frog" respectively. This result suggests that the neural networks in question have a default output for low probability images with respect to distribution of the input domain, which in turn affects adversarial examples computed with large perturbation budgets.

## 7 Conclusion

What was the main idea.

## 8 Contribution Statement

This is joint work from Maximilian Samsinger and Matthias Dellago. Max wrote the code... We thank Alexander Schlögl for the research idea.

## References


[1] 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.

[2] 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.

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

[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.