




Adversarial Label Flips

Matthias Dellago & Maximilian Samsinger

Fast gradient sign method



The diagram illustrates the Fast Gradient Sign Method (FGSM). It shows a sequence of three images: a panda, a noisy gradient image, and a gibbon. The panda image is on the left, followed by a plus sign, a small epsilon symbol, a noisy gradient image, an equals sign, and the gibbon image on the right. Below the panda image is the text "Panda (57.7% confidence)". Below the noisy gradient image is the text $\text{sign}(\nabla_x J(\theta, x, y))$. Below the gibbon image is the text "Gibbon (99.3% confidence)".

Panda
(57.7% confidence)

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

Gibbon
(99.3% confidence)

[2] Explaining and harnessing adversarial examples, 2014

What we want to do

Confusion Matrix

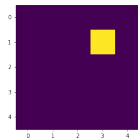
		Categorised as		
		Dog	Cat	Plane
Adversarial Example of a	Dog	0.0	?	?
	Cat	?	0.0	?
	Plane	?	?	0.0

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

Quantifying Changes

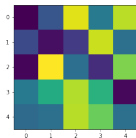
L^0 -Norm

Number of
pixels changed



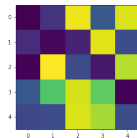
L^1 -Norm

Sum of
all changes



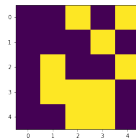
L^2 -Norm

Sum of the *square*
of all changes



L^∞ -Norm

Maximum of
all changes



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.