Adversarial examples

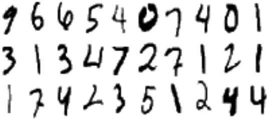
Jan-Hendrik Plank, Jonas Dehning & Philipp Höhne

July 12, 2017

Table of Contents

The data sets

MNIST



Dogs vs. Cats

Networks

- CNN

Data set	Network structure	# Layers
MNIST	5 CL - 2 FC	7
Dogs vs. Cats	11 CL - 3 FC	14

Finding adversarial examples: Gradient Method

$$\vec{\eta} = \epsilon \cdot \mathrm{sign}\left(\nabla_{\vec{x}} J_{loss}\big|_{\vec{x}}\right)$$

$\vec{\eta}$: noise

$\vec{x}$: picture

$\epsilon \ll 1$

Finding adversarial examples: Minimizer

Minimize over η:

$$\min_{\vec{\eta}} \left(\frac{1}{1 + \delta - p(\vec{x} + \vec{\eta})} + c \cdot \sigma(\vec{\eta}) \right)$$

$\vec{\eta}$: noise

$\vec{x}$: picture

p : prediction

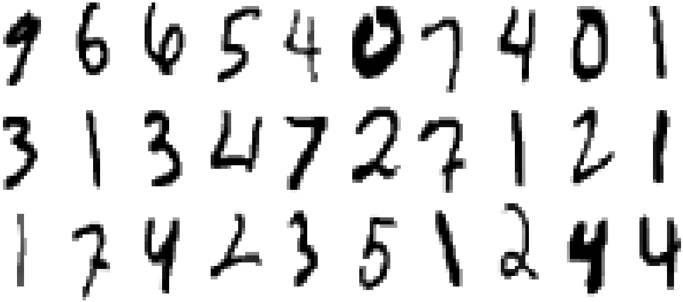
c : a constant

$\delta \ll 1$

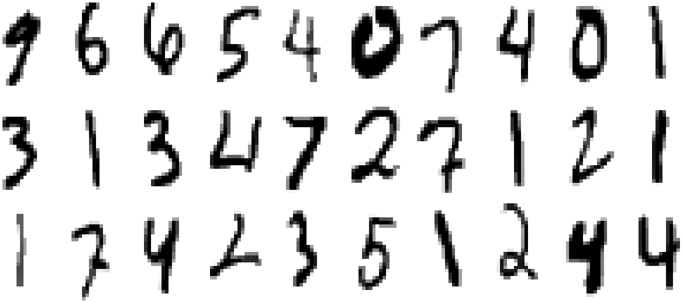
with constraint:

$$\vec{x} + \vec{\eta} \in [0, 1]^n$$

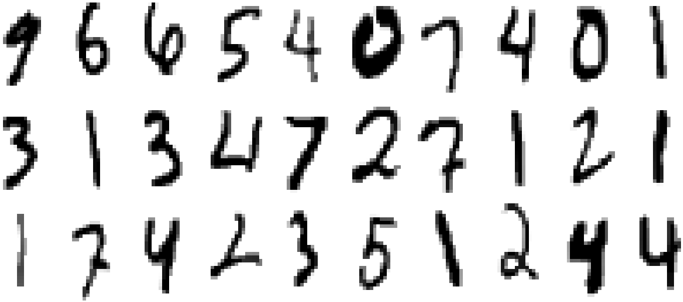
MNIST adversarial examples I



MNIST adversarial examples II



Dogs vs. Cats adversarial examples I



Dogs vs. Cats adversarial examples II