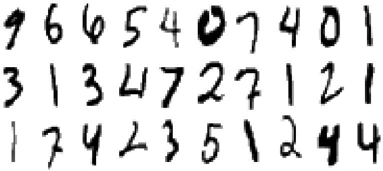
Digit recognition and adversarial examples

Jan-Hendrik Plank, Jonas Dehning & Philipp Höhne

June 1, 2017

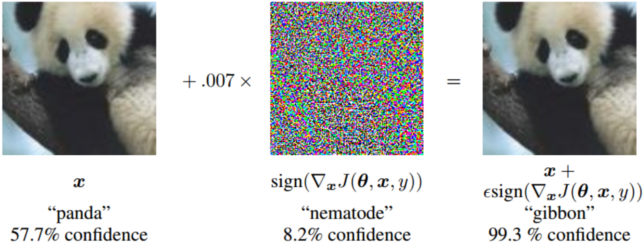
Digit recognition



- Method: deep convolution network
- MNIST-Dataset
- 28000 images: 28px × 28px

Adversarial examples

- Small pertubations → false classification



+ .007 ×
=
$\boldsymbol{x}$
"panda"
57.7% confidence
$\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$
"nematode"
8.2% confidence
$\boldsymbol{x} + \epsilon\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$
"gibbon"
99.3 % confidence

- From Goodfellow et al. 2015

Further ideas

- improve network by training with the perturbed images
  - improved kaggle score?
- compare different datasets
- compare robustness of different networks