Adversarial Examples in Machine Learning

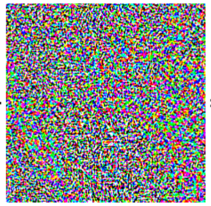
Matěj Kocián
Supervisor: Mgr. Martin Pilát, Ph.D.
Charles University
Faculty of Mathematics and Physics

Introduction

Deep learning has been recently achieving great results in many tasks such as image recognition, machine translation, etc. However, deep neural nets are susceptible to adversarial examples [1] – subtle input perturbations a human observer might not even notice, whereas the neural network gives completely different response:

Original image, recognized as "panda"

+


Sign of the gradient of the error function w.r.t. the input (then multiplied by small epsilon)

=

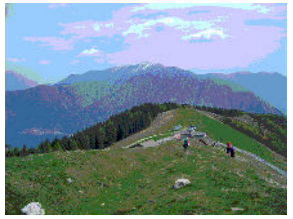
Adversarial example, recognized as "nematode" [2]

Adversarial examples pose a threat to real world systems utilizing machine learning models, such as self-driving cars or authentication systems, especially because they can be transfered from one neural network to another [1] and are effective even with the distortion caused by viewing them from different angles [3, 4]. Although adversarial examples are an active research topic, the problem has not been solved so far. In this work, we propose two new defense methods and evaluate them on standard datasets.

Rounding

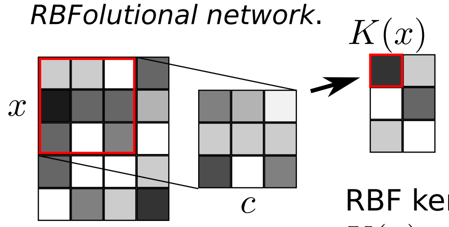
Methods generating adversarial examples depend on the ability to find a point in the input space close to the original input, that will be however classified differently by the network. We tried to discretize the input space per coordinate so as to disable such small perturbations.

Input, possibly adversarially perturbed


Rounded input

RBFolutional network

Our second approach follows from the hypothesis suggesting networks with RBF kernels should not be so easily fooled [2]. However, such networks generally achieve lower accuracy than deep convolutional networks in image classification tasks. We tried combining both approaches into a *RBFolutional network*.



Convolutional "kernel"

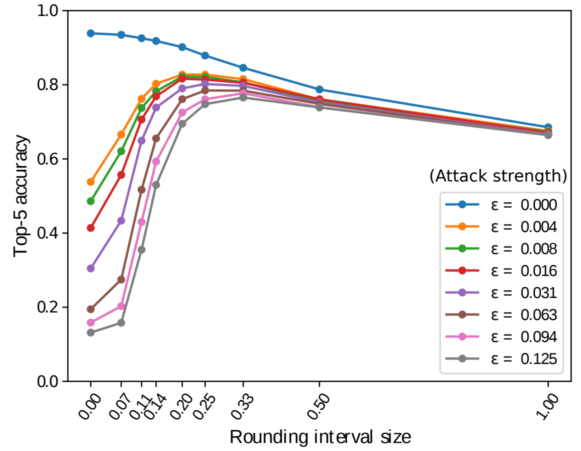
$$K(x) = x \cdot c$$

RBF kernel

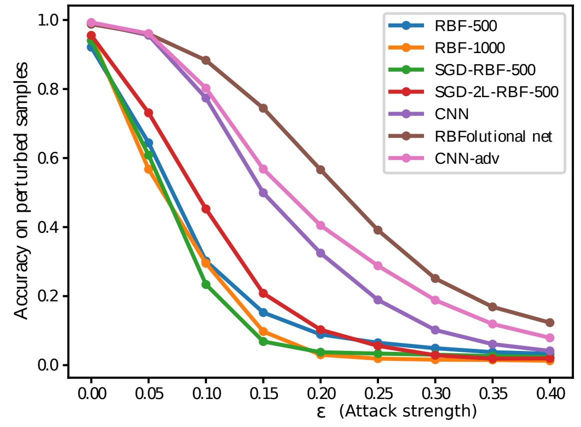
$$K(x) = \exp\left(-\beta \|x - c\|_2^2\right)$$

Results

Rounding was able to reduce adversarial examples effectiveness, especially of those generated by iterative FGSM [3]. Results on ImageNet:



RBFolutional net outperformed both RBF nets and CNNs when evaluated on adversarial examples (FGSM [2]) generated from MNIST images:



[1] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. J. Goodfellow, and R. Fergus. Intriguing properties of neural networks. CoRR, abs/1312.6199, 2013. [2] I. J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. CoRR, abs/1412.6572, 2014. [3] A. Kurakin, I. J. Goodfellow, and S. Bengio. Adversarial examples in the physical world. CoRR, abs/1607.02533, 2016. [4] A. Athalye and I. Sutskever. Synthesizing robust adversarial examples. CoRR, abs/1707.07397, 2017.

matej.kocian@gmail.com