One Pixel Attack for Fooling Deep Neural Networks [1]

Presented by:
Ana Letícia Garcez Vicente

[1] Su, J; Vargas, D. V and Sakurai, K. **One Pixel Attack for Fooling Deep Neural Networks**. October 17, 2019. arXiv: 1710.08864.

Introduction

- Adversarial images
  - small perturbation: one pixel attack
    - Differential Evolution (DE)
  - three and five pixel attack

- Black-box DNN attack
  - only information: probability labels

- Datasets:
  - Original and Kaggle CIFAR-10 (size of 32x32):
    - AllConv, NiN and VGG
    - 0 to 9 indicates, respectively, the classes: airplane, automobile, bird, cat, deer, dog, frog, horse, ship and truck
  - ImageNet (resolutions to 227x227):
    - BVLC AlexNet

One-pixel attacks

Figure 1: ImageNet dataset. Black (original class labels), blue (target class labels) and their confidence. [1]

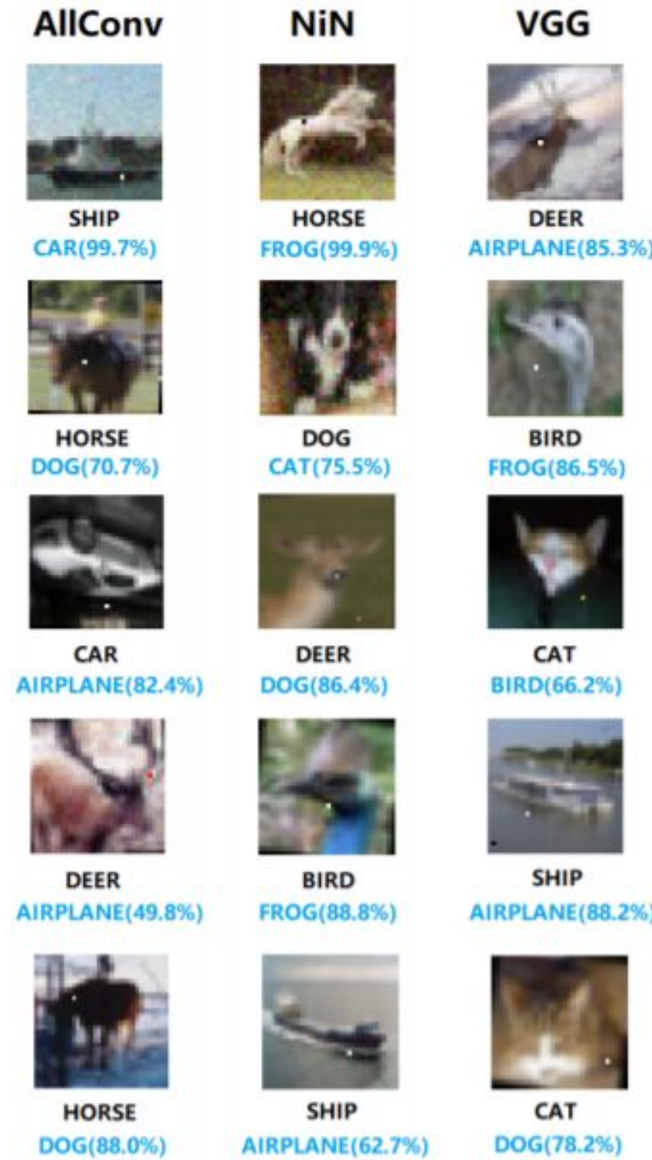


Figure 2: CIFAR-10 dataset. Black (original class labels) and blue (target class labels and the confidence). [1]

Methodology

The process of generating adversarial imagens can be described by an optimization problem with constraints.

Let f be the classifier that receives the input $\boldsymbol{x} = (x_1, \ldots , x_n)$, which is the original natural image classified as classt. Where $f_t(\boldsymbol{x})$ is the probability of $\boldsymbol{x}$ belongs to the classt. And the vector $e(\boldsymbol{x}) = (e_1, \ldots , e_n)$ is the perturbation.

The goal is:

$$\begin{aligned} \underset{e(\mathbf{x})^*}{\text{maximize}} \quad & f_{adv}(\mathbf{x} + e(\mathbf{x})) \\ \text{subject to} \quad & \|e(\mathbf{x})\|_0 \leq d, \end{aligned}$$

where, the *adv* is the target class and *L* is the maximum modification (one-pixel attack *L* = 1). [1]

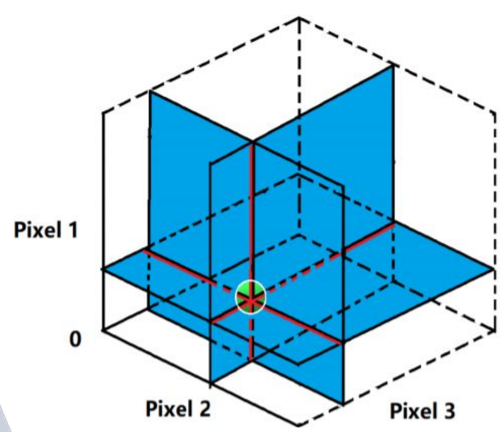


Figure 3: 3-dimensional input space. [1]

Differential Evolution (DE)

- Evolutionary algorithms (EA)
  - keeping diversity
  - improving fitness values
- Advantages: Higher probability of finding global optima, Require less information and Simplicity

Method and Settings

- Each perturbation is a tuple with 5 elements: the coordinates which indicates the position of the pixel and RGB values.
- Initial number candidate solution: 400

$$x_i(g+1) = x_{r1}(g) + F(x_{r2}(g) - x_{r3}(g)),$$
$$r1 \neq r2 \neq r3,$$

where, x_i is the element of candidate solution, r1, r2 e r3 are random numbers, *F* is the scale parameter and g is the current generation. [1]

Effectiveness*

- Success Rate and Adversarial Probability Labels (Confidence):

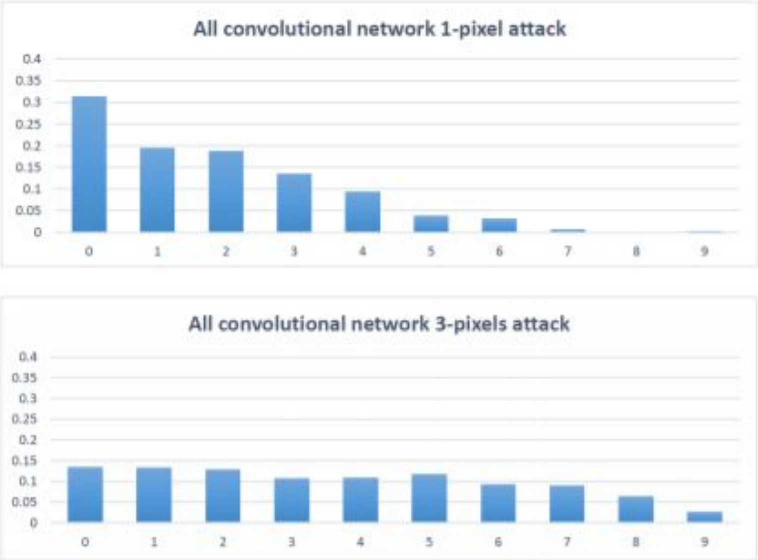
	AllConv	NiN	VGG16	BVLC
OriginAcc	85.6%	87.2%	83.3%	57.3%
Targeted	19.82%	23.15%	16.48%	–
Non-targeted	68.71%	71.66%	63.53%	16.04%
Confidence	79.40%	75.02%	67.67%	22.91%

Table 1: One-Pixel Attack. OriginAcc is the accuracy on the natural test dataset and Target/Non-Target is the accuracy of the attack[1]

	3 pixels	5 pixels
Success rate(tar)	40.57%	44.00%
Success rate(non-tar)	86.53%	86.34%
Rate/Labels	79.17%	77.09%

Table 2: Three-Pixel Attack on AllConv and Five-Pixel Attack on NiN. [1]

- Number of Target Class:



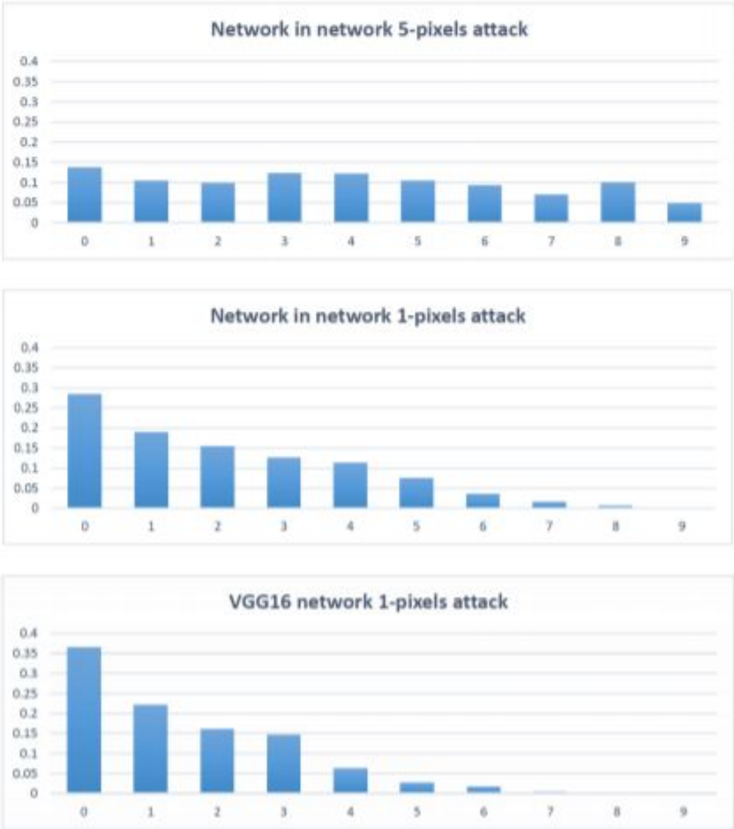


Figure 4: Percentual (vertical axis) of natural images perturbed to a certain number (0 to 9). [1]

*ImageNet dataset is only used by BVLC network

Effectiveness

- Original-Target Class Pairs:

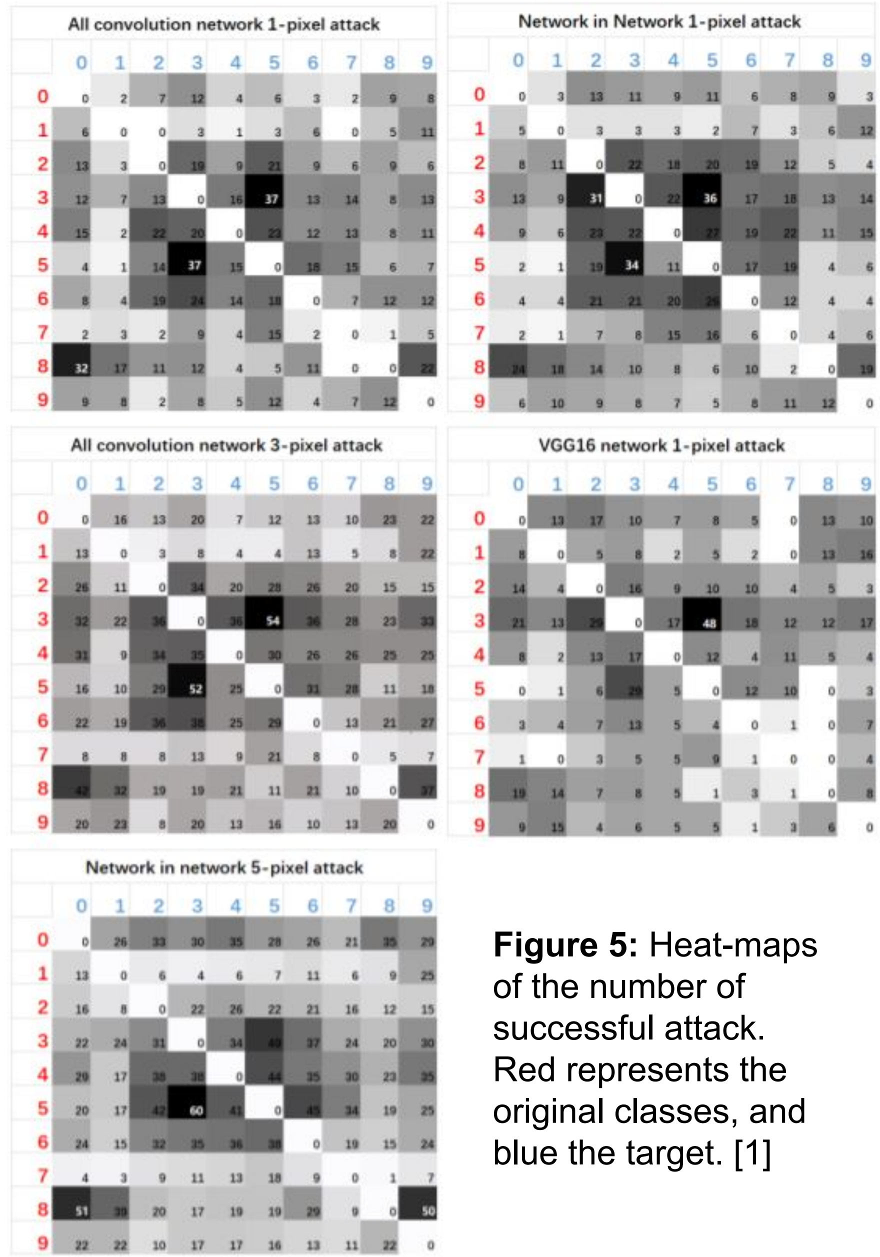


Figure 5: Heat-maps of the number of successful attack. Red represents the original classes, and blue the target. [1]

Effectiveness

- Time Complexity and Average Distortion:

	AllConv	NiN	VGG16	BVLC
AvgEvaluation	16000	12400	20000	25600
AvgDistortion	123	133	145	158

Table 3: Cost of conducting one-pixel attack. AvgEvaluation is the the average number of evaluation to produce adv images. AvgDistortion is the required average distortion to produce adv images[1]

- Random One-Pixel Attack:

	AllConv	NiN	VGG16
DE success rate	68.71%	71.66%	63.53%
Confidence	79.40%	75.02%	67.67%
Random Search success rate	49.70%	41.72%	15.57%
Confidence	87.73%	75.83%	59.90%

Table 4: Non-Targeted attack on Kaggle CIFAR-10 dataset. [1]

- Change in fitness value:

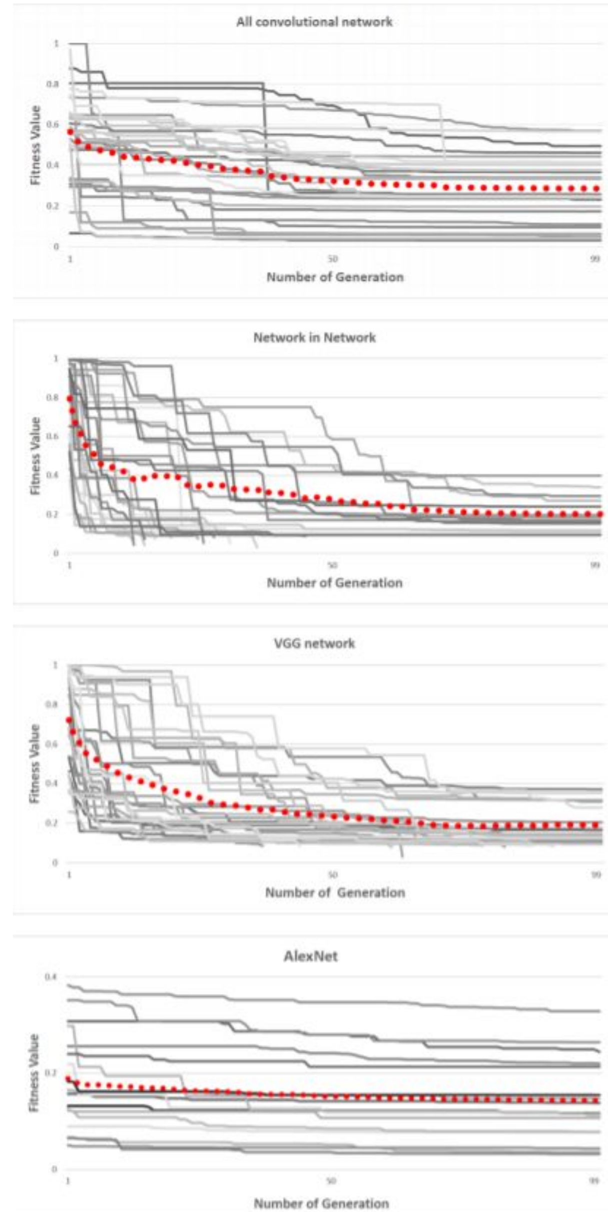


Figure 6: Change of fitness value (100 generations of non-targeted attack). The average is the red dotted lines. [1]

Original CIFAR-10 dataset

- Attack Rate:

	AllConv	NiN	VGG16
Targeted	3.41%	4.78%	5.63%
Non-targeted1	22.67%	32.00%	30.33%
Confidence	54.58%	55.18%	51.19%
Non-targeted2	22.60%	35.20%	31.40%
Confidence	56.57%	60.08%	53.58%

Table 4: One-Pixel Attack. Non-targeted1 is the non-targeted attack accuracy calculated by targeted attack results. Non-targeted2 is the true non-targeted attack accurancy[1]

- Number of Targeted Classes:

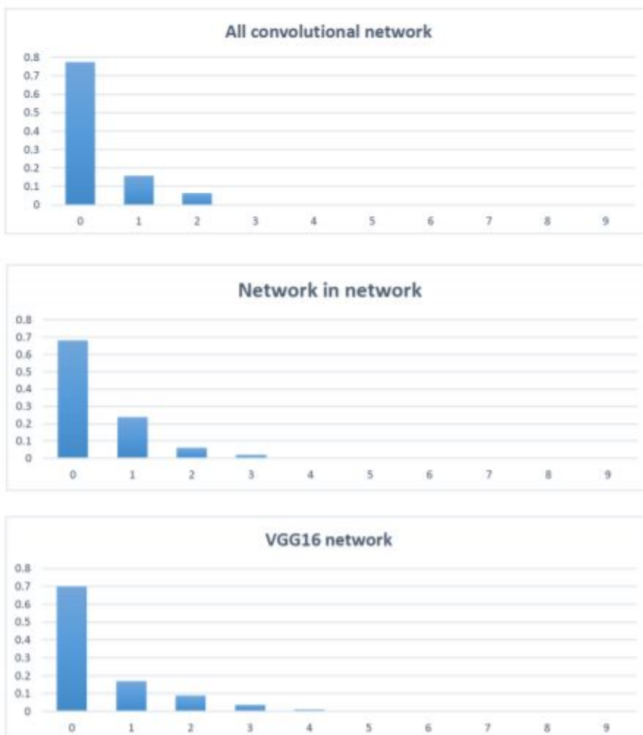


Figure 7: Percentual (vertical axis) of natural images perturbed to a certain number (one-pixel targeted attack). [1]

Original CIFAR-10 dataset

- Original-Target Class Pairs:

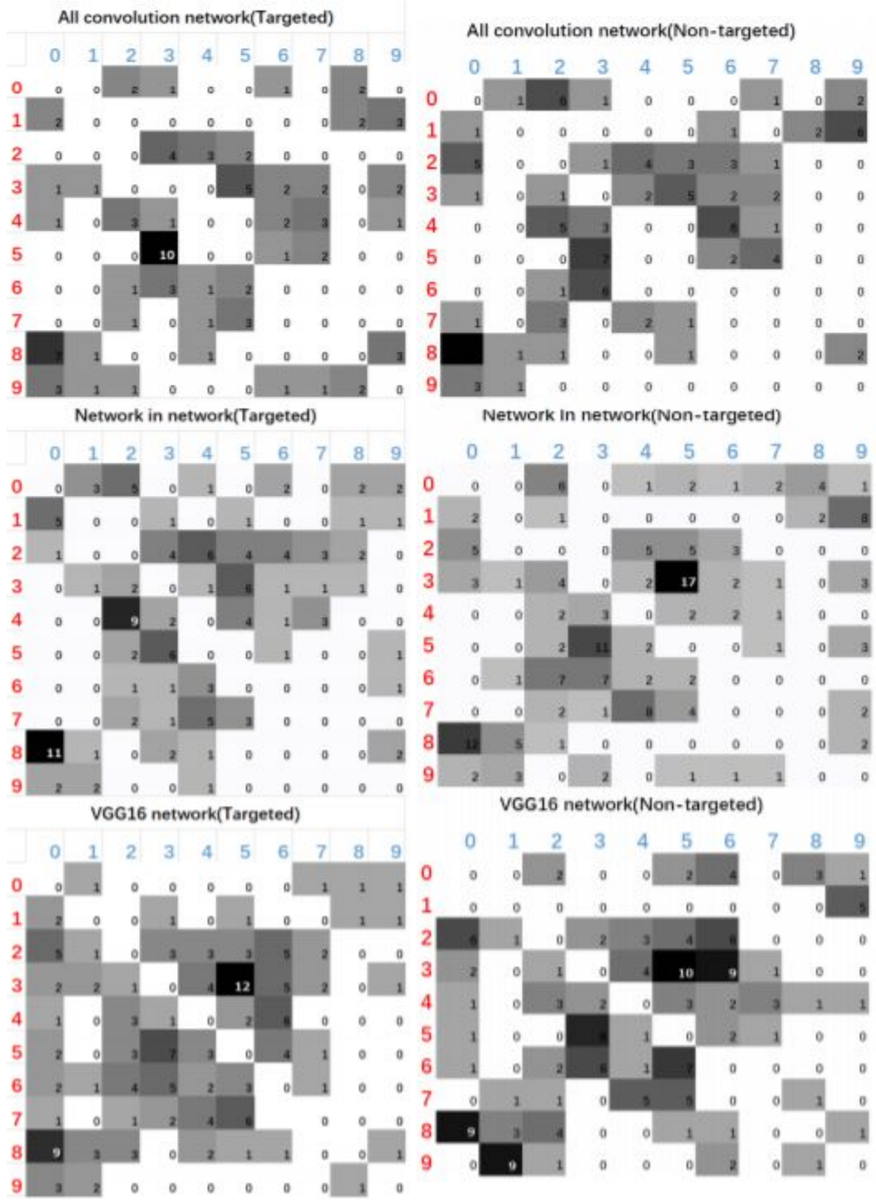


Figure 8: Heat-maps of the number of successful targeted and non-targeted attack. Red represents the original classes, and blue the target.[1]

Summary

- Fundamental problem: neural networks aren't able to ignore the adversarial perturbation
- One-pixel attack is sufficient to fool the network, even with big dimension image (ImageNet)

Next Steps

- Understand more about the boundary of the images
- Evolutionary strategies can improve the method by allowing more efficient and accurate attacks
- Neuroevolution which allows to learn the weights and the network's topology
- Unified neuron model that can adapt the structure to the problem
- Adversarial Training