May 7th, 11:00 AM - 1:00 PM

Exploring and Expanding the One-Pixel Attack

Umairullah Khan
Portland State University

Walt Woods
Portland State University

Christof Teuscher
Portland State University

Exploring and Expanding the One-Pixel Attack

Umair Khan, Walt Woods, Christof Teuscher
Department of Electrical and Computer Engineering

Maseeh College of Engineering and Computer Science
PORTLAND STATE UNIVERSITY

teuscher.:Lab
teuscher-lab.com

1. Introduction

In machine learning research, adversarial examples are normal inputs to a classifier that have been specifically perturbed to cause the model to misclassify the input. Recent work has demonstrated that several image-classifying deep neural networks (DNNs) can be reliably fooled with the modification of single pixel in the input image -- a technique referred to as the "one-pixel attack".

We present data on three avenues of exploration into the attack and consider future research directions:

(i) a modification in technique which produces lower attack RMSE
(ii) a comparison of the attack across different networks
(iii) an analysis of the attack via the generation of per-pixel heatmaps for input images

2. Convolutional Neural Networks (CNNs)

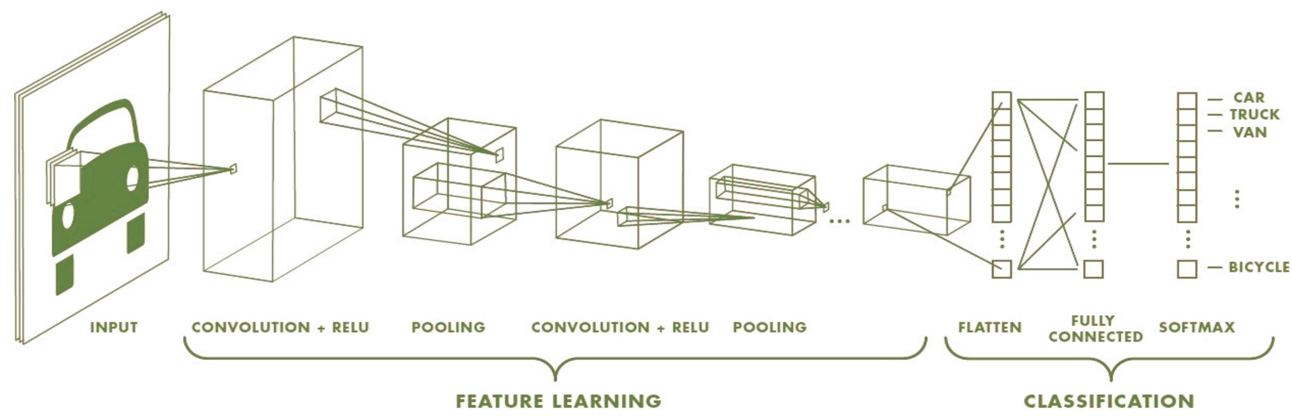

Figure 1 -- The basic architecture of a convolutional neural network.
[Image by Raghav Prabhu, medium.com/@RaghavPrabhu. Colors have been modified.]

- Commonly used for image classification tasks.
- Build on basic neural networks by arranging neurons in three "spatial" dimensions -- height, width, and depth.
- Each grouping of neurons in the same three-dimensional "space" represents a convolution -- in this context, a mathematical operation on a matrix using another matrix (kernel) that results in a new "feature map".
- Through training, these convolutional layers can "learn" features that determine the classification of an image (e.g. eyes, tires, etc).

3. The One-Pixel Attack

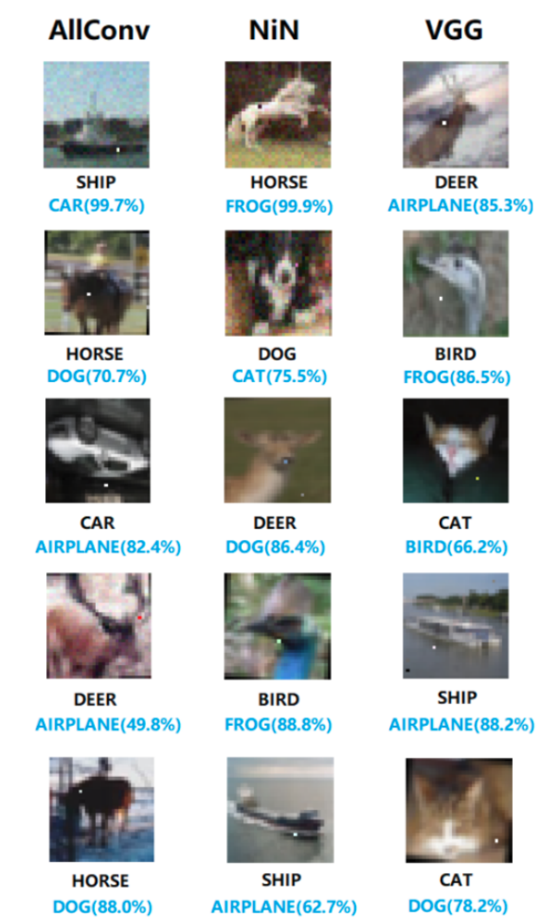

Figure 2 -- A demonstration of the attack.
[Image from [1].]

- Technique initially described by J. Su, D.V. Vargas, and K. Sakurai in 2017 [1].
- Causes convolutional neural networks to misclassify input images by perturbing just one pixel in the input image (Figure 2).
- Perturbations are encoded as five-element vectors (x, y, R, G, B) where the first two elements denote position and the last three encode a color value.
- Attacks are generated using a genetic algorithm known as differential evolution (Figure 3), though crossover is omitted.

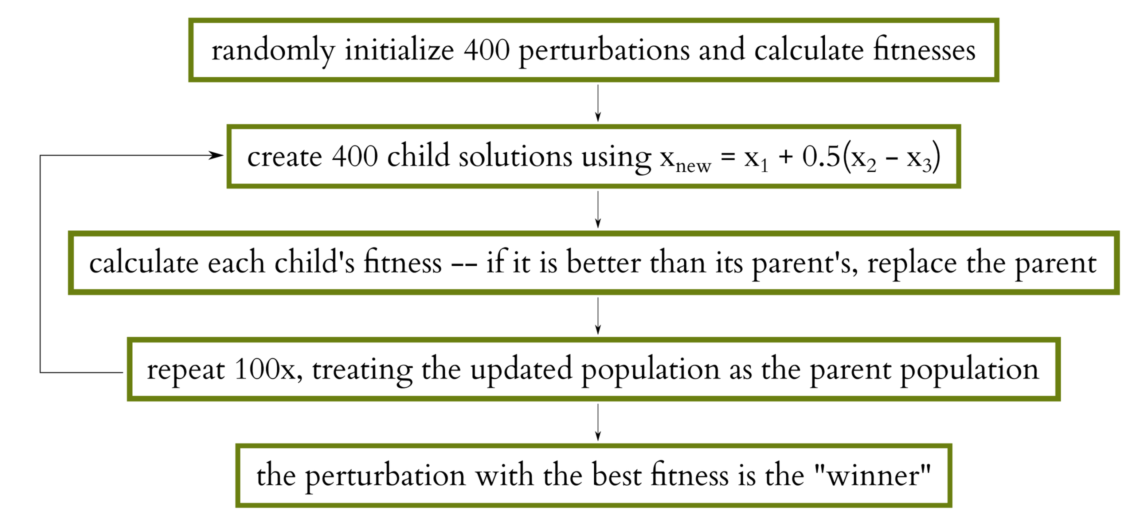

Figure 3 -- A flowchart describing differential evolution. Fitness is defined as the confidence in the correct label; x_1, x_2, and x_3 are randomly selected members of the parent population.

4. Results

(i)

	Success Rate	Mean Attack RMSE
original	37% (181/500)	0.02418
improved	38% (190/500)	0.01946
		19.5% improvement

Table 1 -- A summary of the attack improvement based on the new fitness function. The mean attack RMSE (averaged across 500 images) dropped by 19.5% while the success rate improved slightly.

(ii)

	# of conv. layers	CIFAR-10 Accuracy	Attack Success Rate	Mean Attack RMSE
Basic	2	62%	***38% (190/500)***	0.01946
ResNet8	8	79%	38% (189/500)	***0.01509***
AllConv	9	***91%***	33% (164/500)	0.01808
ResNet14	***14***	87%	28% (139/500)	0.01657

Table 2 -- A summary of the attack performance across CNNs of varying depth. Both the success rate and the mean attack RMSE decreased as networks deepened. Interestingly, RMSE is significantly lower for networks with a residual architecture compared to those without.

(iii)

Figure 4 -- Examples of per-pixel heatmaps on various images (Basic network). Four general cases are described along two axes: successful/unsuccessful attack and effective/ineffective pixel. Examples of all cases are seen in all networks.

(i)

- A common metric used to quantify the "effect" of an adversarial attack is the *root-mean-squared error* (RMSE).
- In this context, defined as: $RMSE = \sqrt{\frac{\sum_{n=1}^{3072}(original - new)^2}{3072}}$
- Flatten the 32x32x3 image into 1x3072, and take the average squared difference between the original and new pixel -- since only one pixel is changed here, only three differences are summed, one for each color channel.
- The original fitness function was simply the confidence of the network in the correct label [1], which did not take RMSE into account -- we revised the function to do so:

$$fitness = \begin{cases} \frac{1}{confidence} & \text{if correctly classified} \\ \frac{1}{RMSE} + 10 & \text{if incorrectly classified} \end{cases}$$

(ii)

- The original paper attacks several networks [1], but all of them are of similar depth.
- To see how the attack performed against networks of varying depth, we implemented and attacked four different convolutional networks ranging from 2-14 layers deep (Table 2).
- Basic: very simple two-convolution network.
- ResNet8: eight-layer implementation of the ResNet architecture [2], which utilizes "residual blocks" to increase accuracy.
- AllConv: essentially an upscaled version of Basic [3].
- ResNet14: 14-layer version of ResNet architecture.
- Though Basic is most vulnerable, other networks have significanly lower average RMSEs.

(iii)

- To analyze the basis of the one-pixel attack and cross-check the performance of DE, we generated per-pixel heatmaps of input images (Figure 4).
- At each pixel, 64 different color perturbations were applied and the maximum confidence change recorded on the heatmap.
- The majority of images are sensitive in at least one region.
- Four conclusions informed by the four rows of Figure 4:
  (1) Some images are vulnerable in many areas. (first row)
  (2) Some images need very little perturbation. (second row)
  (3) Differential evolution does not always find the most effective color. (third row)
  (4) Differential evolution does not always find the most effective pixel to attack. (fourth row)

5. Conclusions

- First and foremost, the data verifies the validity of the one-pixel technique -- convolutional networks are susceptible to attack with minimal perturbation.
- The characteristics of a network have a significant impact on the attack, both in terms of attack success rate and RMSE.
- Differential evolution is clearly not maximizing the potential of the one-pixel theory -- there are many cases where per-pixel analysis shows a high attack potential but differential evolution fails to optimize to the best solution.
- With this in mind, there are a couple of avenues for future exploration:

(1) Can we further improve the success of the one-pixel attack by investgating when and why differential evolution fails and addressing those problems? Or is a new search algorithm necessary?

(2) How can we apply the one-pixel attack to other domains where machine learning is used (e.g. video applications)?

6. Acknowledgements / References

The authors acknowledge the support of the Maseeh College of Engineering and Computer Science (MCECS) through the Undergraduate Research and Mentoring Program (URMP).

Contact: {ukhan, wwoods, teuscher}@pdx.edu

[1] J. Su, D.V. Vargas, and K. Sakurai. "One pixel attack for fooling deep neural networks", arXiv:1710.08864v5 [cs.LG], Jan. 2019.
[2] K. He, X. Zhang, S. Ren, and J. Sun. "Deep Residual Learning for Image Recognition", presented at the *IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, Jun. 2016.
[3] J.T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller. "Striving for Simplicity: The All Convolutional Net", arXiv:1412.6806v3 [cs.LG], Apr. 2015.