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*

Let us know how access to this document benefits you.

Follow this and additional works at: https://pdxscholar.library.pdx.edu/studentsymposium

Part of the Computer Sciences Commons, and the Engineering Commons

Khan, Umairullah; Woods, Walt; and Teuscher, Christof, "Exploring and Expanding the One-Pixel Attack" (2019). *Student Research Symposium*. 2.

https://pdxscholar.library.pdx.edu/studentsymposium/2019/Posters/2

This Event is brought to you for free and open access. It has been accepted for inclusion in Student Research Symposium by an authorized administrator of PDXScholar. For more information, please contact pdxscholar@pdx.edu.
Exploring and Expanding the One-Pixel Attack

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

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 a 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)

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.

<table>
<thead>
<tr>
<th>CNN</th>
<th>Original Success Rate</th>
<th>Original Mean Attack RMSE</th>
<th>Improved Success Rate</th>
<th>Improved Mean Attack RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Basic</td>
<td>38% (196/500)</td>
<td>0.01946</td>
<td>38% (196/500)</td>
<td>0.01909</td>
</tr>
<tr>
<td>ResNet8</td>
<td>7% (34/500)</td>
<td>0.01946</td>
<td>8% (40/500)</td>
<td>0.01897</td>
</tr>
<tr>
<td>AllConv</td>
<td>93% (460/500)</td>
<td>0.01808</td>
<td>93% (460/500)</td>
<td>0.01808</td>
</tr>
<tr>
<td>ResNet14</td>
<td>14% (69/500)</td>
<td>0.01657</td>
<td>14% (69/500)</td>
<td>0.01657</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>CNN</th>
<th># of conv. layers</th>
<th>CIFAR-10 Accuracy</th>
<th>Attack Success Rate</th>
<th>Mean Attack RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Basic</td>
<td>2</td>
<td>62%</td>
<td>38% (196/500)</td>
<td>0.01946</td>
</tr>
<tr>
<td>ResNet8</td>
<td>8</td>
<td>79%</td>
<td>8% (40/500)</td>
<td>0.01509</td>
</tr>
<tr>
<td>AllConv</td>
<td>9</td>
<td>91%</td>
<td>33% (164/500)</td>
<td>0.01808</td>
</tr>
<tr>
<td>ResNet14</td>
<td>14</td>
<td>87%</td>
<td>26% (130/500)</td>
<td>0.01657</td>
</tr>
</tbody>
</table>

Figure 1 — The basic architecture of a convolutional neural network.

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

3. The One-Pixel Attack

- Technique initially described by J. Su, D.V. Vargas, and K. Sakurai in 2017 [1].
- Causes convolutional neural networks to misclassify inputs 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.

Table 3 — A summary of the attack performance across CNNs of varying depth. Both the success rate and the attack RMSE decreased as networks deepened. Interestingly, RMSE is significantly lower for networks with a residual architecture compared to those without.

<table>
<thead>
<tr>
<th>CNN</th>
<th>Original Success Rate</th>
<th>Original Mean Attack RMSE</th>
<th>Improved Success Rate</th>
<th>Improved Mean Attack RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Basic</td>
<td>38% (196/500)</td>
<td>0.01946</td>
<td>38% (196/500)</td>
<td>0.01909</td>
</tr>
<tr>
<td>ResNet8</td>
<td>7% (34/500)</td>
<td>0.01946</td>
<td>8% (40/500)</td>
<td>0.01897</td>
</tr>
<tr>
<td>AllConv</td>
<td>93% (460/500)</td>
<td>0.01808</td>
<td>93% (460/500)</td>
<td>0.01808</td>
</tr>
<tr>
<td>ResNet14</td>
<td>14% (69/500)</td>
<td>0.01657</td>
<td>14% (69/500)</td>
<td>0.01657</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>CNN</th>
<th># of conv. layers</th>
<th>CIFAR-10 Accuracy</th>
<th>Attack Success Rate</th>
<th>Mean Attack RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Basic</td>
<td>2</td>
<td>62%</td>
<td>38% (196/500)</td>
<td>0.01946</td>
</tr>
<tr>
<td>ResNet8</td>
<td>8</td>
<td>79%</td>
<td>8% (40/500)</td>
<td>0.01509</td>
</tr>
<tr>
<td>AllConv</td>
<td>9</td>
<td>91%</td>
<td>33% (164/500)</td>
<td>0.01808</td>
</tr>
<tr>
<td>ResNet14</td>
<td>14</td>
<td>87%</td>
<td>26% (130/500)</td>
<td>0.01657</td>
</tr>
</tbody>
</table>

Figure 3 — A demonstration of the attack. [Image from [1]].

4. Results

- A common metric used to quantify the "effect" of an adversarial attack is the root-mean-squared error (RMSE).
- In this context, defined as: \( \text{RMSE} = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i|^2 \)\.
- 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: \( \text{fitness} = \begin{cases} \text{count classified} & \text{if correctly classified} \\ \text{count misclassified} + 10 & \text{if incorrectly classified} \end{cases} \)\.
- 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).\n- 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 significantly lower average RMSEs.
- The original paper attacks several networks [1], but all of them are of similar depth.

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 technique -- 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 investigating when and why differential evolution fails and addressing those problems?
  (2) Or is a new search algorithm necessary?
  (3) How can we apply the one-pixel attack to other domains where machine learning is used (e.g. video applications)?
  (4) Di\’erential evolution does not always find the most effective color. (third row)
  (5) Differential evolution does not always find the most effective pixel to attack. (fourth row)

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