One Pixel Attack for Fooling Deep Neural Networks

Aash Mohammad
2022PDS0021@iitjammu.ac.in

Bibek Mondal
2022PDS0022@iitjammu.ac.in

Rishiraj Mukati
2022PIS0028@iitjammu.ac.in

May 18, 2023

1 Abstract

Recent studies have uncovered the susceptibility of Deep Neural Networks (DNNs) to manipulation through small perturbations applied to input vectors. This project focuses on analyzing a specific attack scenario, where only a single pixel can be modified. To address this, we propose a novel approach utilizing differential evolution (DE) to generate one-pixel adversarial perturbations. Our method operates as a black-box attack, requiring minimal adversarial information, and demonstrates the ability to deceive a wider range of network types due to the inherent characteristics of DE. Our results reveal that, 67.97% of the natural images in the Kaggle CIFAR-10 test dataset can be perturbed to target at least one different class by modifying just one pixel, achieving an average confidence of 74.03%. This attack sheds light on the vulnerability of current DNNs to low-dimensional attacks, presenting a distinctive perspective in the field of adversarial machine learning within an extremely constrained scenario.

2 Introduction

In the field of image recognition, Deep Neural Network (DNN)-based approaches have surpassed traditional image processing techniques, achieving performance comparable to human-level accuracy. However, recent studies have exposed the vulnerability of DNNs to artificial perturbations applied to natural images, resulting in misclassifications. Various effective algorithms, known as "adversarial image" generation methods, have been proposed to create such perturbed samples. These methods typically involve adding a small and carefully crafted additive perturbation to a correctly classified natural image, aiming to be imperceptible to human observers. This slight modification can cause the classifier to misclassify the perturbed image into a completely different class.

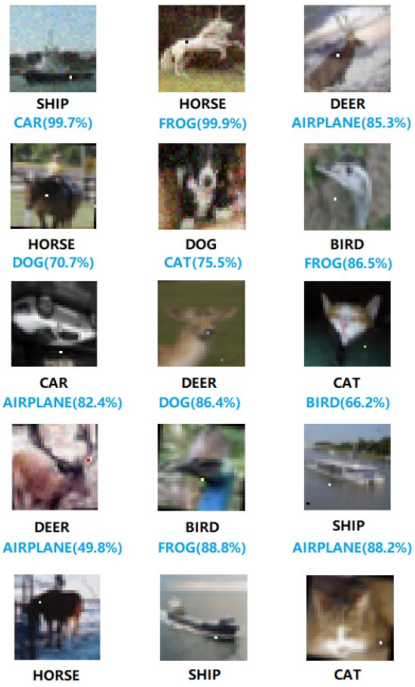

Figure 1: One Pixel attack fooled DNN trained on CIFAR-10 datasets (The original class labels are in black color while the target class labels and the corresponding confidence are given below.)

In this project, we focus on the generation of adversarial images using a one-pixel perturbation approach based on differential evolution. This black-box attack operates in a scenario where only the probability labels of the DNN are accessible, without any detailed knowledge of its internal workings. By leveraging differential evolution, we perturb a single pixel to deceive the classifier and induce misclassifications.

3 Methodology

3.1 Problem Description

Generating adversarial images can be formulated as an optimization problem with constraints. We represent an input image as a vector, where each scalar element corresponds to a pixel. Let f be the target image classifier that takes n-dimensional inputs, and let $\mathbf{x} = (x_1, ..., x_n)$ be the original natural image correctly classified as class t. The probability of $\mathbf{x}$ belonging to class t is denoted as $f_t(\mathbf{x})$.

We introduce the vector $\mathbf{e}(\mathbf{x}) = (e_1, ..., e_n)$ as an additive adversarial perturbation applied to $\mathbf{x}$, with the target class being adv and a maximum modification limit denoted as L. The length of vector $\mathbf{e}(\mathbf{x})$ measures the extent of modification, and our objective as adversaries in targeted attacks is to find the optimized solution $\mathbf{e}(\mathbf{x})^*$ that maximizes the following expression:

$$\max_{\mathbf{e}(\mathbf{x})^*} f_{adv}(\mathbf{x} + \mathbf{e}(\mathbf{x})^*)$$

subject to the constraint $||\mathbf{e}(\mathbf{x})^*|| \leq L$.

In our approach, we slightly modify the equation as follows:

$$\max_{\mathbf{e}(\mathbf{x})^*} f_{adv}(\mathbf{x} + \mathbf{e}(\mathbf{x})^*)$$

subject to the constraint $||\mathbf{e}(\mathbf{x})^*||_0 \leq d$, where d is a small number. In the case of a one-pixel attack, $d = 1$ only d dimensions are perturbed while the other dimensions of $\mathbf{e}(\mathbf{x})$ remain zero.

The one-pixel modification can be seen as perturbing the data point along a direction parallel to one of the n dimensions.

3.2 Differential Evolution

Differential Evolution (DE) is a population-based optimization algorithm widely used for solving complex multi-modal optimization problems. It falls under the category of evolutionary algorithms and incorporates mechanisms in the population selection phase to maintain diversity, allowing it to efficiently find high-quality solutions compared to gradient-based or other evolutionary algorithms. DE generates a set of candidate solutions (children) based on the current population (parents) during each iteration. These children are then compared to their corresponding parents, with only the fittest ones surviving. This process simultaneously promotes diversity and improves fitness values.

DE offers several advantages when it comes to generating adversarial images:

- **Higher probability of Finding Global Optima:** DE has a greater likelihood of finding global optima compared to gradient descent or greedy search algorithms. This advantage stems from its diversity-maintaining mechanisms and the use of a set of candidate solutions.
- **Require Less Information from Target System:** DE does not rely on gradient information, making it applicable to a wider range of optimization problems, including non-differentiable, dynamic, or noisy scenarios. This characteristic is particularly valuable for generating adversarial images, as calculating gradients often necessitates extensive information about the target system, which may not be readily available in many cases.
- **Simplicity:** Another strength of DE is its simplicity and independence from the specific classifier used. The proposed attack method can be applied as long as the probability labels of the target system are known, without requiring detailed knowledge of the underlying classifier.

By utilizing DE for generating adversarial images, these advantages contribute to the effectiveness and practicality of the approach.

3.3 Method and Settings

We encode the perturbation into an array, which serves as a candidate solution for optimization using differential evolution (DE). Each candidate solution contains a fixed number of perturbations, where each perturbation is represented by a tuple consisting of the x-y coordinates and RGB value. Each perturbation modifies a single pixel in the image.

Initially, the population size is set to 10. In each iteration, an additional 10 candidate solutions (children) are generated using the DE formula:

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

where x_i represents an element of the candidate solution, $r1$, $r2$, and $r3$ are random numbers, F is the scale parameter set to 0.5, and g is the current generation index.

After generation, each candidate solution competes with its corresponding parent based on the population index. Only the fittest candidate survives for the next iteration. The maximum number of iterations is set to 100. For targeted attacks on the Kaggle CIFAR-10 dataset, the early-stop criterion is triggered when the probability label of the target class exceeds 90%.

4 Results

4.1 Classes Affected

The analysis of the results presented in Figure 2 reveals that a considerable proportion of natural images can be manipulated to exhibit misclassifications into two, three, or even four different target classes with just a minimal modification of a single pixel. Moreover, it is observed that as the number of modified pixels is increased, the probability of achieving perturbations that lead to misclassifications into a larger number of target classes becomes significantly higher. These findings emphasize the vulnerability of deep neural networks to adversarial attacks, even in scenarios where only a limited number of pixels are altered.

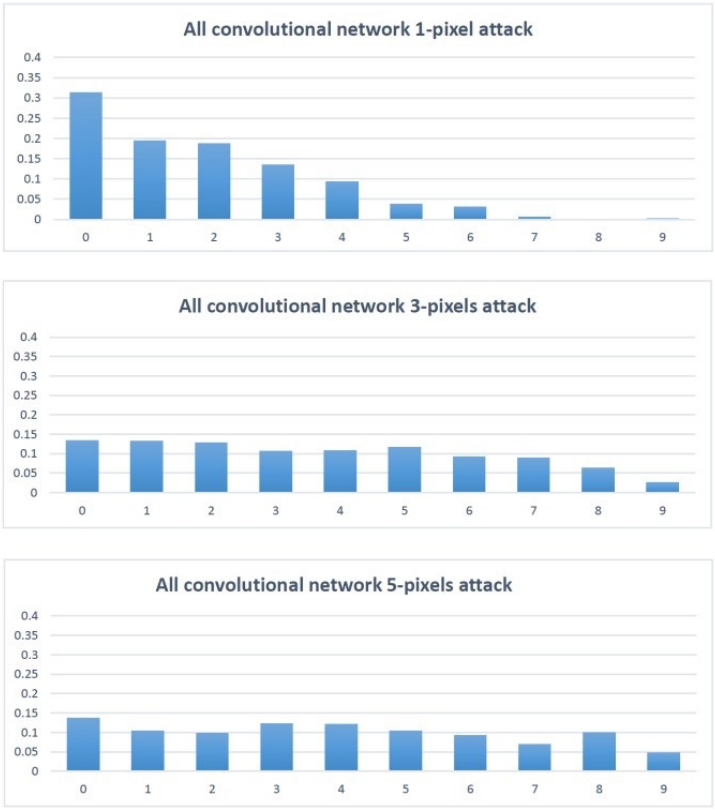

Figure 2: The graphs shows the percentage of natural images that were successfully perturbed to a certain number (from 0 to 9) of target classes by using one, three or five-pixel perturbation. The vertical axis shows the percentage of images that can be perturbed while the horizontal axis indicates the number of target classes.

4.2 Original-MissClassified Class Pairs

Figure 3 highlights that certain combinations of original and missclassified classes are more susceptible to perturbations than others. For instance, images classified as cats (class 3) are more easily manipulated to be misclassified as dogs (class 5), while achieving a misclassification to automobiles (class 1) is considerably more challenging. This suggests that vulnerable target classes share common directions in the input space, which are applicable to multiple data points belonging to the same class.

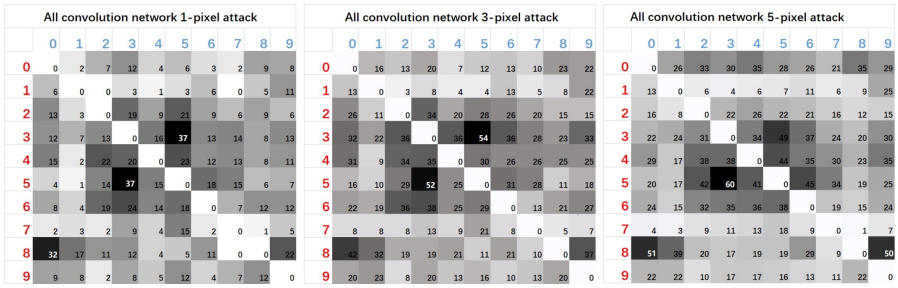

Figure 3: The graphs shows the percentage of natural images that were successfully perturbed to a certain number (from 0 to 9) of target classes by using one, three or five-pixel perturbation. The vertical axis shows the percentage of images that can be perturbed while the horizontal axis indicates the number of target classes.

5 Discussion

5.1 Robustness of One-pixel Attack

The robustness of the one-pixel attack depends on several factors, including the specific machine learning model being targeted, the dataset it was trained on, and the defense mechanisms in place. Here are some considerations regarding the robustness of the one-pixel attack:

- **Model architecture:** Different models have varying sensitivities to perturbations in input images. Some models may be more robust to the one-pixel attack due to their inherent architecture, while others may be more vulnerable. The attack is more likely to succeed against models that heavily rely on pixel-level features for classification.
- **Dataset characteristics:** The robustness of the attack can be influenced by the diversity and quality of the dataset on which the targeted model was trained. If the training dataset contains similar images with minimal variations in pixel values, the attack may be more effective. However, if the dataset is diverse and contains variations in pixel values, the attack's success rate may decrease.
- **Attack parameters:** The success of the one-pixel attack depends on the choice of the pixel to modify and the desired target class. Different combinations of attack parameters can yield varying results. By exploring different pixel positions and target classes, an attacker can increase the likelihood of a successful attack.
- **Defense mechanisms:** Various defense techniques can be employed to mitigate the effectiveness of the one-pixel attack. Some common approaches include adversarial training, input pre-processing, and model assembling. These defenses can make the models more robust by detecting or reducing the impact of adversarial examples.
- **Transferability:** The transferability of the attack refers to its effectiveness across different models or instances of the same model. If the attack is successful against a particular model, it might also be successful against other similar models. Transferability can enhance the practical impact of the one-pixel attack.

Overall, the one-pixel attack can be a powerful technique to exploit vulnerabilities in machine learning models. However, the robustness of the attack depends on various factors, and its effectiveness may vary across different models and datasets. Ongoing research is focused on developing more robust models and defenses to counter adversarial attacks like the one-pixel attack.

6 References

1. J. Brest, S. Greiner, B. Boskovic, M. Mernik, and V. Zumer. Self-adapting control parameters in differential evolution: A comparative study on numerical benchmark problems. *IEEE transactions on evolutionary computation*, 10(6), pp. 646-657, 2006.

2. S. Das and P. N. Suganthan. Differential evolution: A survey of the state-of-the-art. *IEEE transactions on evolutionary computation*, 15(1), pp. 4-31, 2011.

3. S. M. Moosavi Dezfooli, F. Alhussein, and F. Pascal. Deepfool: a simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2574-2582, 2016.

4. S. M. Moosavi Dezfooli, F. Alhussein, F. Omar, F. Pascal, and S. Stefano. Analysis of universal adversarial perturbations. *arXiv preprint arXiv:1705.09554*, 2017.

5. N. Narodytska and S. Kasiviswanathan. Simple black-box adversarial attacks on deep neural networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1310-1318. IEEE, 2017.

6. N. Papernot, P. McDaniel, I. Goodfellow, S. Jha, Z. B. Celik, and A. Swami. Practical black-box attacks against machine learning. In *Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security*, pp. 506-519. ACM, 2017.