PixelDefend: Leveraging Generative Models to Understand and Defend against Adversarial Examples

Yang Song¹, Taesup Kim², Sebastian Nowozin³, Stefano Ermon¹, Nate Kushman³
¹Stanford University, ²Université de Montréal, ³Microsoft Research

ABSTRACT

• We show that generative models can be used for detecting adversarially perturbed images and observe that most adversarial examples lie in low probability regions.

• We introduce a novel family of methods for defending against adversarial attacks based on the idea of purification.

• We show that a defensive technique from this family, PixelDefend, can achieve state-of-the-art results on a large number of attacking techniques, improving the accuracy against the strongest adversary on the CIFAR-10 dataset from 32% to 70%.

ADVERSARIAL EXAMPLES

State-of-the-art classifiers can be fooled by adding quasi-imperceptible noise.

NEURAL DENSITY MODELS

PixelCNN a convolutional neural network that factorizes \( p(X) \) using the product rule

\[
p(X) = \prod_{i=1}^{n} p(x_i | x_{1:i-1}).
\]

where the pixel dependencies are in raster scan order.

DETECTING ADVERSARIAL EXAMPLES

Observation: The PixelCNN density of an adversarial example is usually significantly lower than that of a clean example. Therefore, \( p(X) \) can be used as a test statistic to detect adversarial examples.

Statistical test: Given an input \( X' \sim q(X) \) and training images \( X_1, X_2, ..., X_N \sim p(X) \). The null hypothesis is \( H_0: p(X) = q(X) \) while the alternative is \( H_1: p(X) \neq q(X) \). The p-value is computed as

\[
p\text{-value} = \frac{1}{N + 1} \left( \sum_{i=1}^{N} I[p(X_i) \leq p(X')] + 1 \right)
\]

EXPERIMENTS

Figure 1: Various attacks of an image from CIFAR-10. The text above shows the attacking methods while the text below shows the predicted labels (of a ResNet).

Figure 2: PixelCNN.

Figure 3: Sampled images for Fashion-MNIST and CIFAR-10. Above red line are real images. Below read line are PixelCNN samples.

Figure 4: (Left) Likelihoods of different adversarial examples. (Right) ROC curves for detecting various attacks.

Figure 5: Distributions of p-values for different attacks.

Figure 6: Adversarial images (left) and purified images after PixelDefend (right).

PIECE DE FEND

Intuition: The harm of adversarial examples might be reduced if they can be modified to have higher likelihood.

Algorithm 1 PixelDefend

Input: Image \( X \), Defense parameter \( \epsilon \), Training images \( X_1, X_2, ..., X_N \)
Output: Purified Image \( X' \)

1. for each row \( i \) do
2. for each column \( j \) do
3. \( \alpha = \epsilon \times \begin{cases} 0 & \text{if } \text{row}(i, j) \text{ is odd} \\ 1 & \text{if } \text{row}(i, j) \text{ is even} \end{cases} \)
4. Set feasible range \( R = \max(\alpha - \epsilon, 0), \min(\alpha + \epsilon, 255) \)
5. Compute the 256-by-software \( \text{pcn}(X) \)
6. Update \( X'[i, j, :N] = \text{arg max}_{N} \text{pcn}(X + \epsilon \times R, i, j) \)
7. end for
8. end for

Table 1: Fashion MNIST (\( \epsilon_{\text{pixels}} = 8/25, \epsilon_{\text{pixels}} = 32 \))

Table 2: CIFAR-10 (\( \epsilon_{\text{pixels}} = 2/25, \epsilon_{\text{pixels}} = 10 \))