Measuring the degree of Robustness of CNN’s towards targeted adversarial examples

Animesh Paul
CSE Graduate Student
University at Buffalo – SUNY
UB Person Number - 50290441

What are Adversarial Examples?

- Inputs to Deep Learning Models that an attacker/hacker intentionally designs to cause the model to misclassify examples.
- Poses a serious threat to mission critical AI systems. As a result, these are also known as adversarial attacks.
- Given an input x and any target classification t (where t is not the label of x), it is possible to find a new input x’ that is similar to x based on a given distance metric but classified as t. x’ is known as a **targeted adversarial example**.

Attempted Defenses against Adversarial Examples

- Adversarial Training:
  A Brute-Force solution wherein a lot of adversarial examples are generated and the model is explicitly trained so as to learn these as fake examples.

- Defensive Distillation (Implemented in Project):
  A strategy where the model is trained to output probabilities of different classes, rather than hard decisions using a modified softmax function. This creates a model whose surface is smoothed in the directions an adversary will typically try to exploit. The level of overfitting is reduced and *blind-spots* are eliminated which an attacker could potentially try to exploit.

An Insight into Defensive Distillation

$$\text{softmax}(x, T)_i = \frac{e^{x_i/T}}{\sum_j e^{x_j/T}}$$

Defensive Distillation proceeds in four steps:

- Train a teacher network using the modified softmax function on hard labels.
- Generate soft labels by applying the teacher network to the training data using modified softmax function.
- Train a distilled network (same shape as the teacher) on soft labels using modified softmax function.
- Finally, upon running the distilled network at test time (to classify new inputs) use T=1.

Defensive distillation successfully defeated traditional attack algorithms and reduced their success probability from 95% to 0.5% [1].

The L-2 Attack

- This is an optimization algorithm with a few constraints.

  -> Minimize L2(x, x + d) such that C(x + d) = t, and C(x) != t where x is fixed and the goal is to find d that minimizes L2(x, x + d).

  -> Here L2(.) measures the standard euclidean distance between the two vectorized images. d is the minimum deviation required to cause the model to misclassify an image.
- Use of an Objective Function:

  -> Because C(x + d) is highly non linear, it has been expressed in another form that is better suited for optimization.

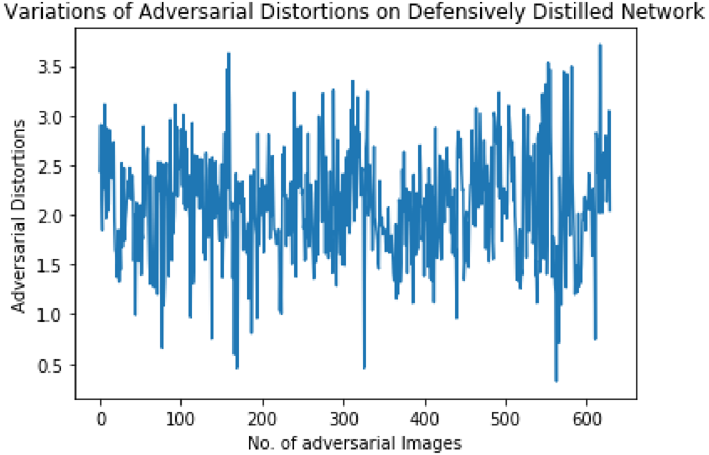
  -> Define an objective function f such that C(x + d) = t if and only if f(x + d) <= 0.

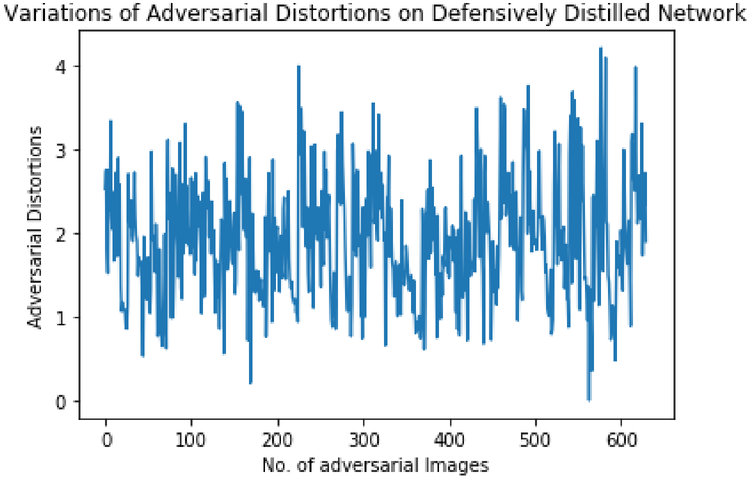
The L-2 Attack (Contd.)

- The optimization function boils down to:
  -> minimize L2(x, x + d) + c * f (x + d) such that x + d is as small as [0, 1]^n.
  -> In other words, L2(||(x + d) - x||) + c * f (x + d) behaves as a loss function while the algorithm attempts to discover the optimal deviation (d).

Observations

- The attack is executed on two convolutional neural networks - undistilled and a defensively distilled neural network.
- Below are variations of distortions(mean adversarial euclidean distances) for both settings. Interestingly, the distances almost remain same in both settings.

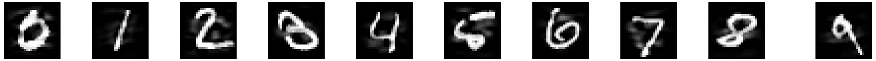




Adversarial Examples Generated

- On Undistilled Network: Below are a few samples of each digit.

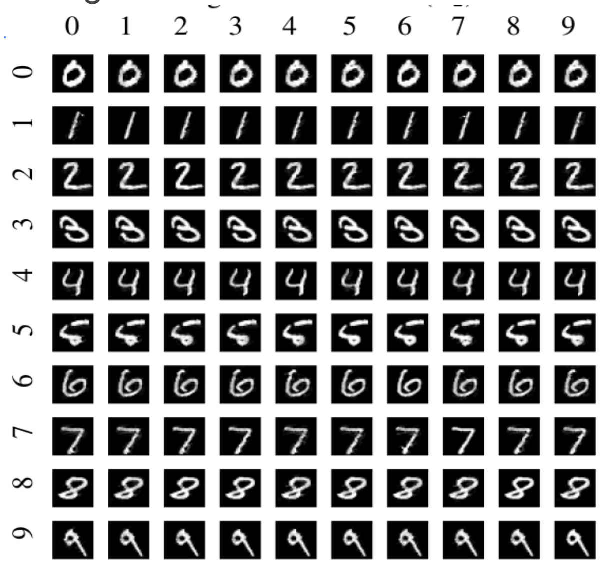
- On Defensively Distilled Network: Below are a few samples of each digit.



- These images have perturbations too minute to be detected by an ordinary neural network as well as a defensively distilled neural network. Even to the human eye, these perturbations are often undetectable.
- A total of 100 images have been generated for each source image corresponding to every targeted label.

Results

- This following figure has been referred from [1] due to difficulties in generating such a grid. Two sets of 100 images have been generated each for each network.
- Each targeted adversarial attack correspond to a source image and a targeted label.
- The images in each row Correspond to the source image with the Digit representing the row.
- The images corresponding to each column Correspond to a misclassified target label as Represented by the column number.
- Undistilled Network: Min L2: 0.007, Avg L2: 1.91
- Defensively Distilled Network: Min L2: 0.32, Avg L2: 2.09



References

1. Towards Evaluating the Robustness of Neural Networks authored by Nicholas Carlini & David Wagner, University of California, Berkeley.
2. https://openai.com/blog/adversarial-example-research/
3. https://towardsdatascience.com/about-adversarial-examples-2a7a7b4d2670