Alter the Structure of Adversarial Perturbations with Model-Agnostic Defense

Yueqiu Sun, Nan Su, Daoyang Shan

Introduction

Deep neural networks are vulnerable to adversarial perturbations and are easily fooled. In our project, we present an evaluation of different defense schemes that alter the structure of adversarial perturbation and defend against adversarial. The defense schemes we investigate is model-agnostic. Therefore, It is difficult for the attacker to circumvent the defense due to the randomness of the method. The evaluation we obtained can be used as reference for deploying different defense strategy under different circumstances.

Adversarial Attack Methodology

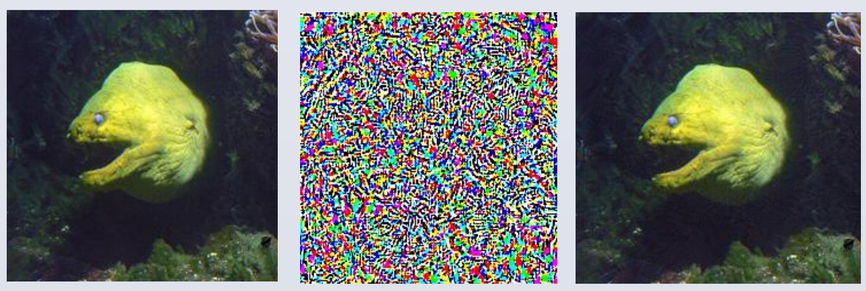
Adversarial examples are imperceptible to human but can easily fool Deep Neural Networks in the testing/deploying stage.

In this project we try three different attack methods:

Fast Gradient Sign Method(FGSM): Performing one step gradient update along the direction of the sign of gradient at each pixel.

Iterative - Fast Gradient Sign Method(I-FGSM): Extending Fast Gradient Sign method by running a finer optimization (smaller change) for multiple iterations. In each iteration, it clips pixel values to avoid large change on each pixel.

L-BFGS: Using a L-BFGS method to minimize the distance between the image and the adversarial as well as the cross-entropy between the predictions for the adversarial and the target class.



Original image
(predicted as eel)
+
Adversarial noise
(added using I-FGSM attack)
+
Generated image
(predicted as bee)

Defense Methodology

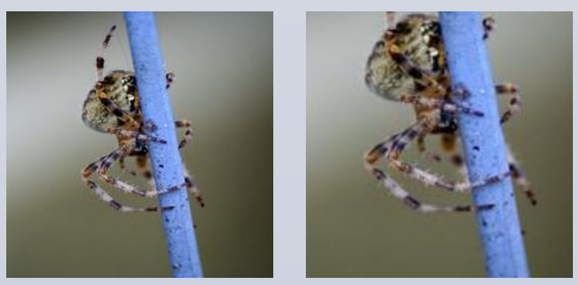
Our assumption on the success of adversarial attack is that such attack produces noise with a specific structure that helps fooling the neural network. To negate this attack, destroying this structure by applying transformation on images before getting the prediction would possibly work. In our project we attempt several transformation methods.

Gaussian Noise: adds random perturbation following Gaussian Distribution on to the initial image.

Speckle Noise: a granular noise that widely exists as image noise.

Gaussian Blur: smooth the initial image with Gaussian Function.

Random Crop: randomly crop the initial image and then resize back.



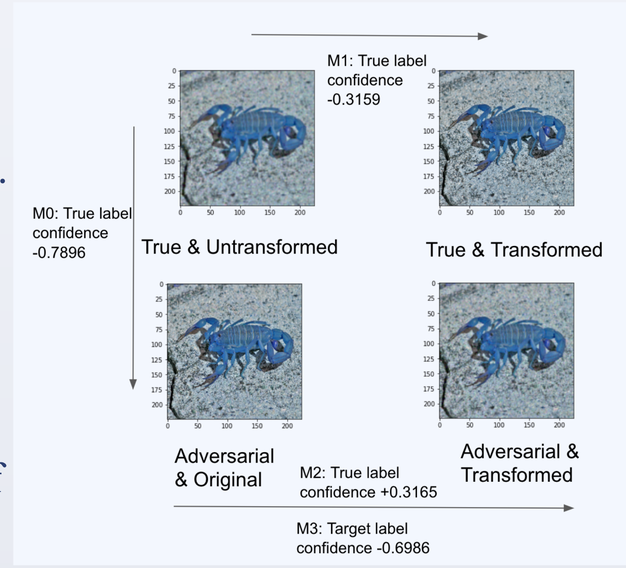
Sample Random Crop: Note the transformed image must have the same size as original image does.

We aim to investigate how applying different combinations of transformation with different parameter as defense schemes on input data influences the classification result of both original and adversarial image. The evaluation of the defense schemes is based on three metrics:

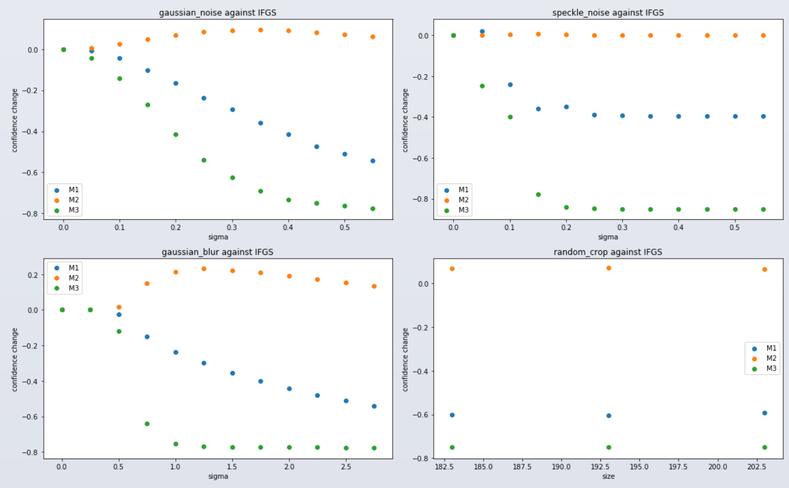
1.The gap between confidence of true class for untransformed sample and transformed sample. (M1)

2.The increase of confidence of true class for adversarial samples.(M2)

3.The decrease of confidence of target class for adversarial samples.(M3)



Results



- The above graph shows the effectiveness of four proposed input transformation to defend against IFGSM.
- Gaussian Blur is the most effective tool among four choices with regard to the metrics while Random Crop is not promising in defending adversarial attack.
- We try some combinations of input transformation to see if they are better than using only one transformation.

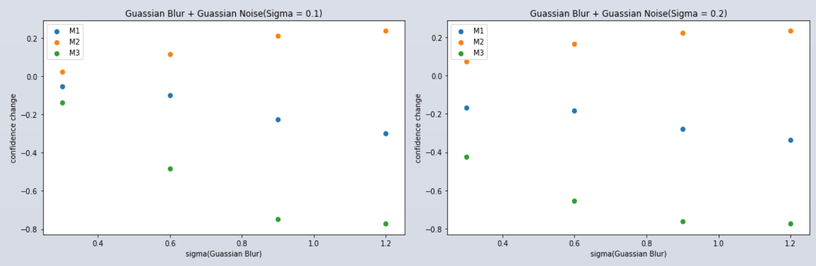


Table 1: Evaluation of the defense schemes against I-FGSM attack

	Gaussian Noise (sigma = 0.25)	Gaussian Blur (sigma = 1)	Speckle Noise (sigma = 0.1)	Random Crop (size = 203)	Gaussian Blur (sigma = 1.2) + Gaussian Noise (sigma = 0.1)	Speckle Noise (sigma = 0.1) + Gaussian Blur (sigma = 0.6)
M1	-0.23641397	-0.23594731	-0.242053	-0.59308326	-0.29859361	-0.11260712
M2	0.08504383	0.21500573	0.00398819	0.06539721	0.2378646	0.11711541
M3	-0.53920394	-0.7541101	-0.39943257	-0.7502144	-0.77034885	-0.48814908

- Gaussian Blur + Gaussian Noise is the most effective in situation where the decrease of confidence in target class(M3) is more valued.
- Speckle Noise + Gaussian Blur is most effective in situation where the tolerance for the gap between confidence of true class for untransformed sample and transformed sample(M1) is small.