

1. Generate 10 adversarial images (1 per class) from each dataset (i.e. MNIST, Fashion- MNIST and CIFAR-10 dataset) by using different attack techniques.

FGSM

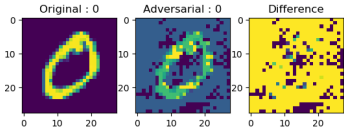
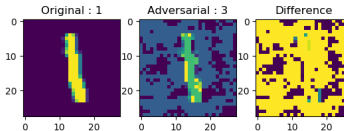
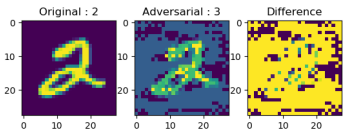
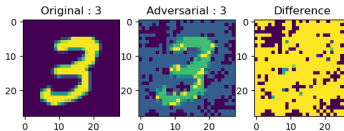
It is to linearize the cost function used to train a model around the neighborhood of the training point. The resulting adversarial example corresponding to the input x is computed as below:

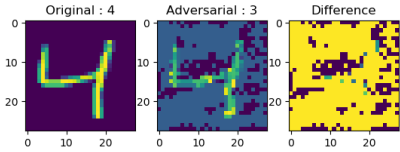
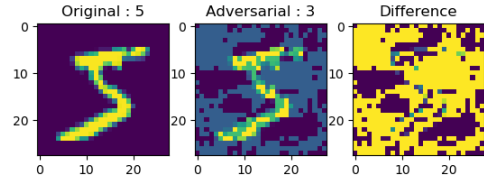
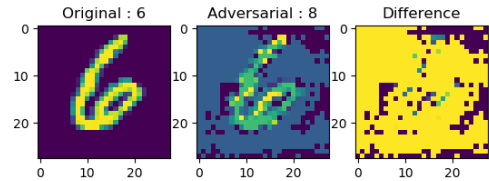
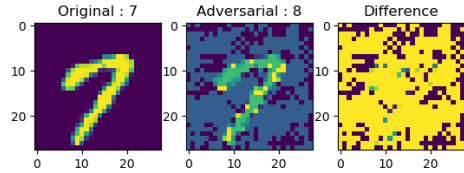
$$x \leftarrow x + \varepsilon \cdot \nabla J(f, \theta, x)$$

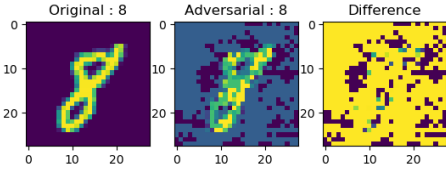
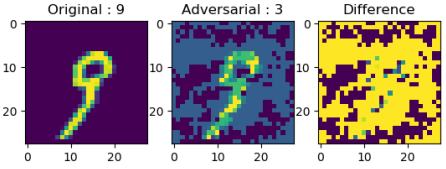
- MNIST

The L2 distance is calculated by

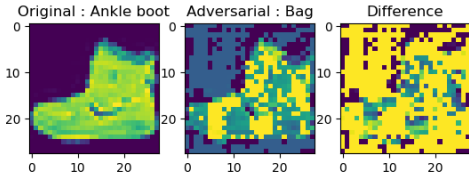
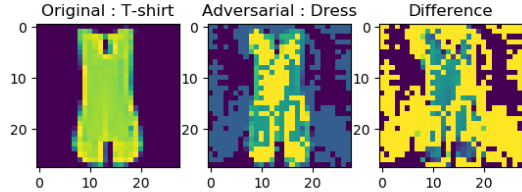
$$d(Adv, Ori) = \sqrt{[\Sigma(Adv[:, :] - Ori[:, :])]^2}$$

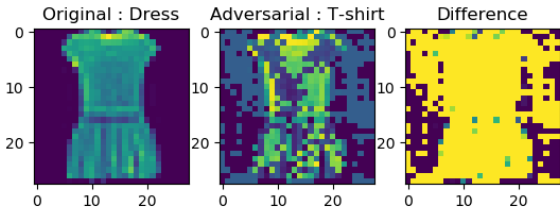
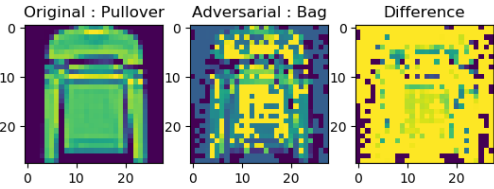
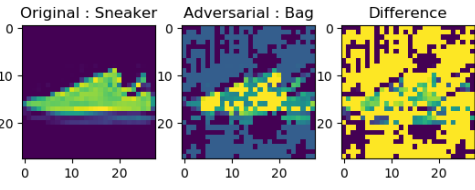
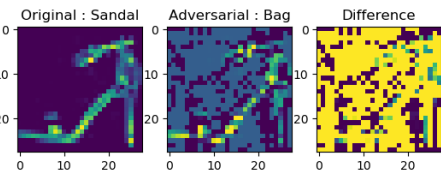
	L2 Distance: 58.638947
	49.684406
	53.834755
	54.681152

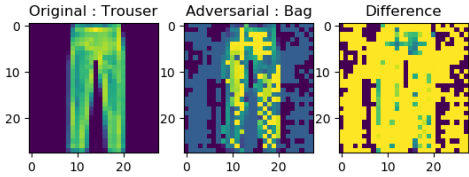
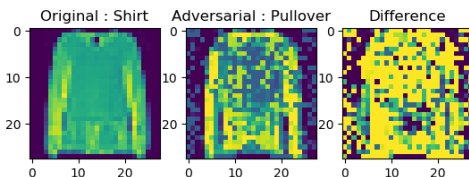
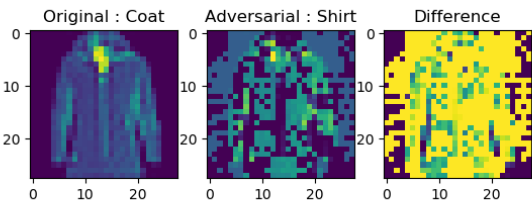
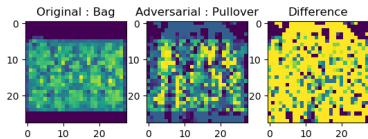
	52.770157
	43.568077
	56.02453
	50.098133

	53.93727
	44.248154

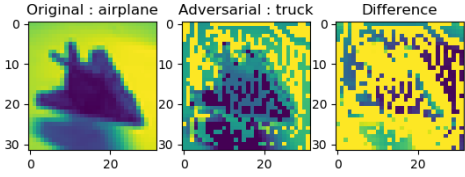
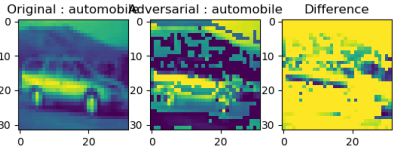
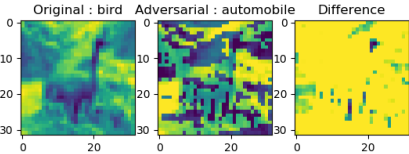
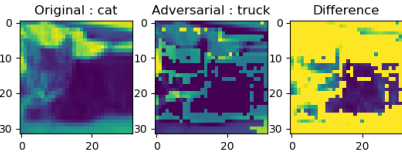
- Fashion MNIST

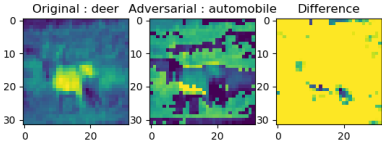
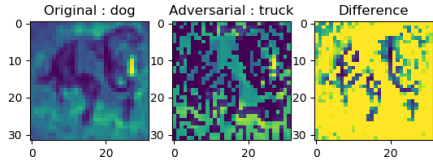
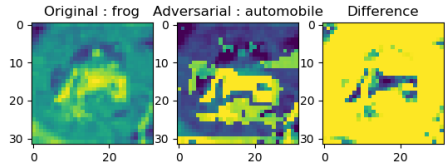
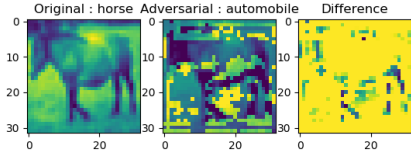
	L2 Distance: 49.731827
	52.33737

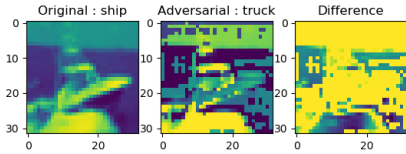
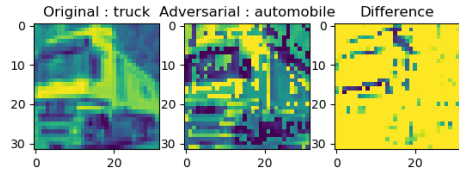
	52.33743
	57.61750
	46.305008
	49.135406

	52.868507
	58.084663
	50.613808
	58.124342

- Cifar10

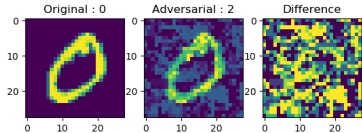
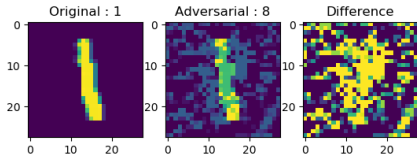
 <p>Original : airplane Adversarial : truck Difference</p>	<p>L2 Distance: 58.34123</p>
 <p>Original : automobile Adversarial : automobile Difference</p>	<p>60.342</p>
 <p>Original : bird Adversarial : automobile Difference</p>	<p>70.34234</p>
 <p>Original : cat Adversarial : truck Difference</p>	<p>40.34237</p>

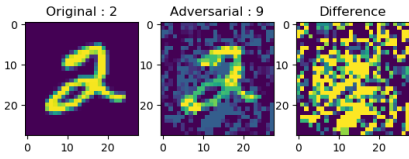
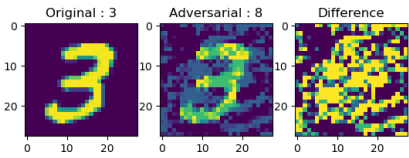
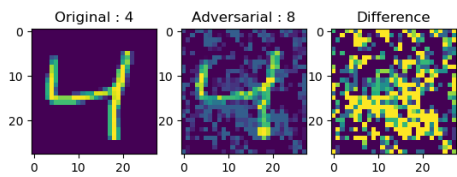
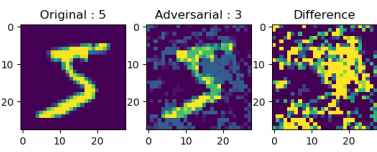
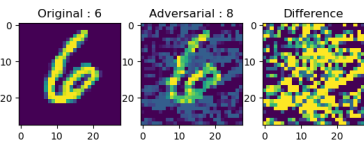
 <p>Original : deer Adversarial : automobile Difference</p>	70.59182
 <p>Original : dog Adversarial : truck Difference</p>	60.23486
 <p>Original : frog Adversarial : automobile Difference</p>	65.45893
 <p>Original : horse Adversarial : automobile Difference</p>	58.65934

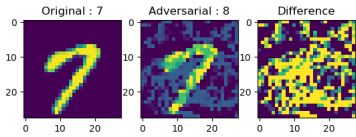
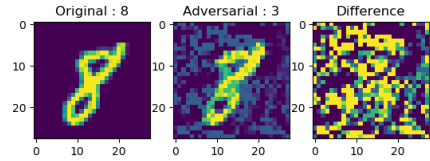
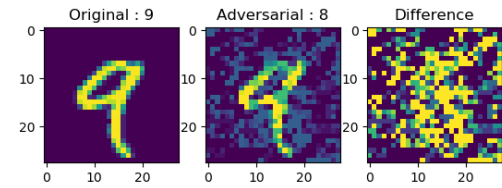
	50.83458
	58.485

Basic Iterative Method

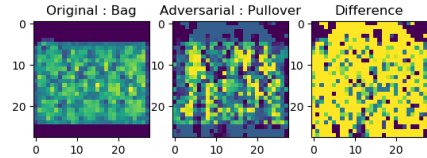
- MNIST

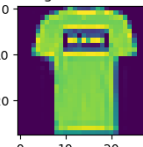
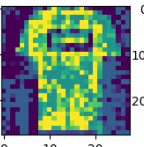
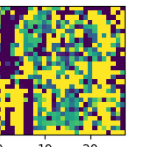
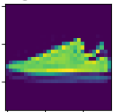
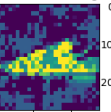
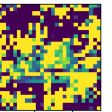
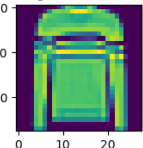
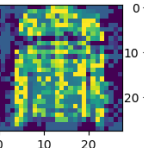
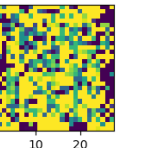
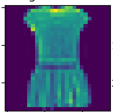
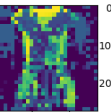
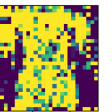
	L2 Distance: 68.2348
	69.3845

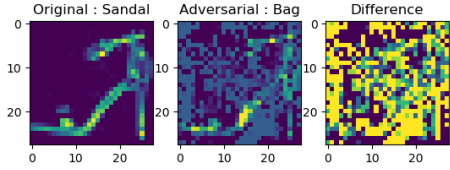
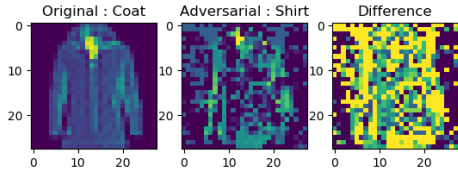
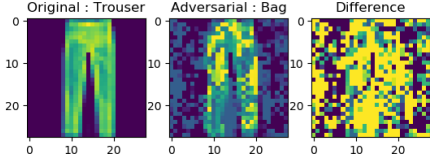
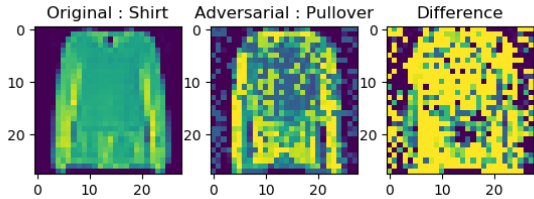
	65.2384
	20.459
	54.148
	44.349
	34.549

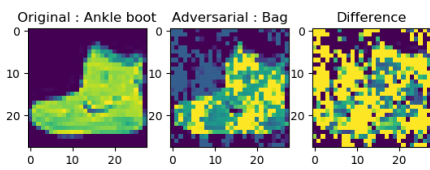
	49.4859
	45.48235
	33.12348

- Fashion MNIST

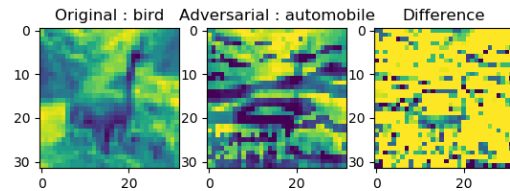
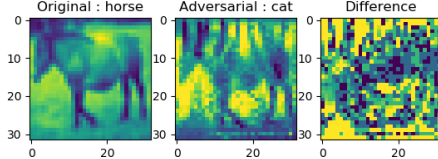
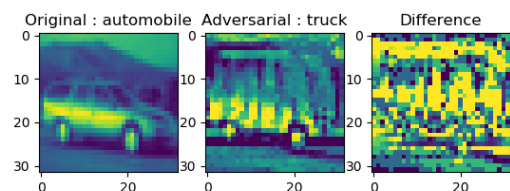
	<p>L2 Distance:</p> <p>36.150993</p>
---	--------------------------------------

<p data-bbox="252 320 718 495">Original : T-shirt  0 10 20</p> <p data-bbox="422 320 566 495">Adversarial : Pullover  0 10 20</p> <p data-bbox="582 320 718 495">Difference  0 10 20</p>	<p data-bbox="1075 197 1216 230">35.113922</p>
<p data-bbox="252 719 630 860">Original : Sneaker  0 10 20</p> <p data-bbox="391 719 502 860">Adversarial : Bag  0 10 20</p> <p data-bbox="518 719 630 860">Difference  0 10 20</p>	<p data-bbox="1075 618 1216 651">32.874474</p>
<p data-bbox="252 1090 710 1256">Original : Pullover  0 10 20</p> <p data-bbox="422 1090 566 1256">Adversarial : Shirt  0 10 20</p> <p data-bbox="582 1090 710 1256">Difference  0 10 20</p>	
<p data-bbox="252 1480 630 1621">Original : Dress  0 10 20</p> <p data-bbox="391 1480 502 1621">Adversarial : T-shirt  0 10 20</p> <p data-bbox="518 1480 630 1621">Difference  0 10 20</p>	<p data-bbox="1075 1379 1216 1413">38.022118</p>

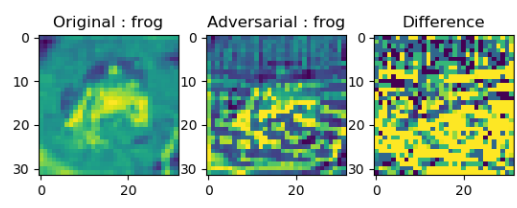
 <p>Original : Sandal Adversarial : Bag Difference</p> <p>The first row displays three heatmaps for the 'Sandal' category. The 'Original' heatmap shows the sandal's shape. The 'Adversarial' heatmap, labeled 'Bag', shows a noisy pattern. The 'Difference' heatmap highlights the changes between the original and adversarial versions.</p>	<p>29.258118</p>
 <p>Original : Coat Adversarial : Shirt Difference</p> <p>The second row displays three heatmaps for the 'Coat' category. The 'Original' heatmap shows the coat's shape. The 'Adversarial' heatmap, labeled 'Shirt', shows a noisy pattern. The 'Difference' heatmap highlights the changes between the original and adversarial versions.</p>	<p>40.22332</p>
 <p>Original : Trouser Adversarial : Bag Difference</p> <p>The third row displays three heatmaps for the 'Trouser' category. The 'Original' heatmap shows the trousers' shape. The 'Adversarial' heatmap, labeled 'Bag', shows a noisy pattern. The 'Difference' heatmap highlights the changes between the original and adversarial versions.</p>	<p>32.052715</p>
 <p>Original : Shirt Adversarial : Pullover Difference</p> <p>The fourth row displays three heatmaps for the 'Shirt' category. The 'Original' heatmap shows the shirt's shape. The 'Adversarial' heatmap, labeled 'Pullover', shows a noisy pattern. The 'Difference' heatmap highlights the changes between the original and adversarial versions.</p>	<p>32.017265</p>

 <p>Original : Ankle boot Adversarial : Bag Difference</p>	<p>33.219147</p>
---	------------------

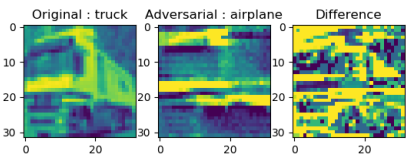
- **Cifar10**

 <p>Original : bird Adversarial : automobile Difference</p>	<p>L2 Distance: 67.3489</p>
 <p>Original : horse Adversarial : cat Difference</p>	<p>65.34823</p>
 <p>Original : automobile Adversarial : truck Difference</p>	<p>79.3458</p>

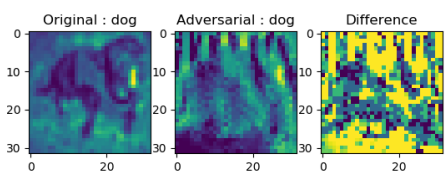
68.34588



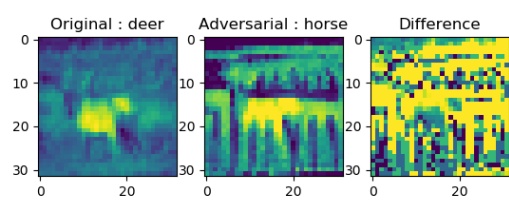
78.3458

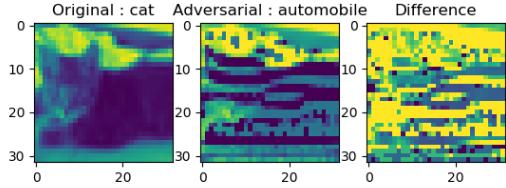
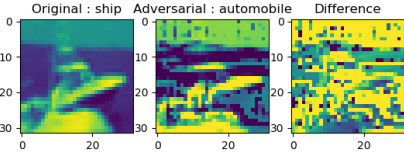
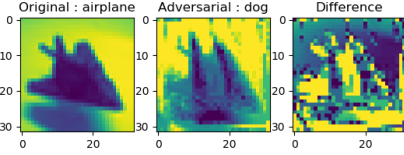


68.3412



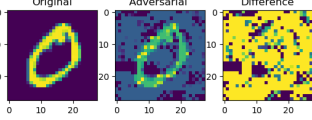
54.2913

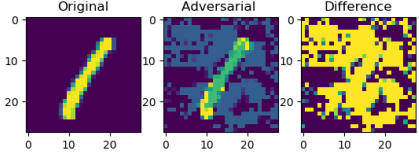
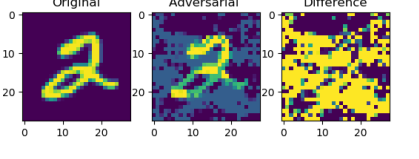
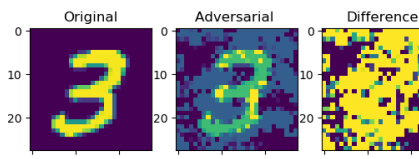
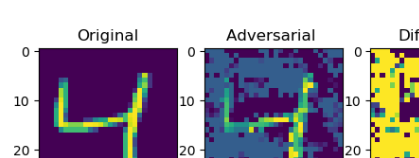


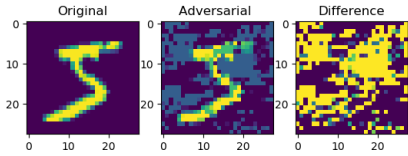
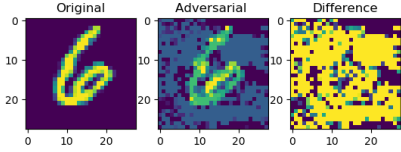
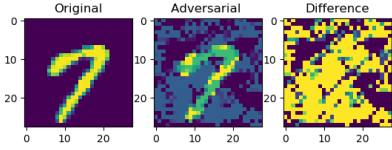
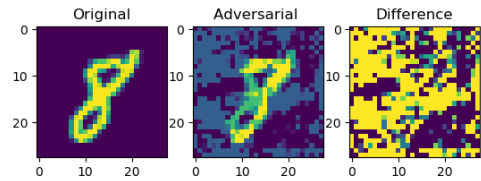
	32.495
	50.1923
	60.4903

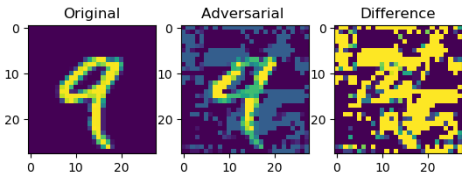
Momentum Iterative method

- MNIST

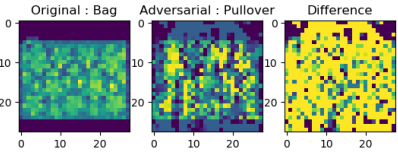
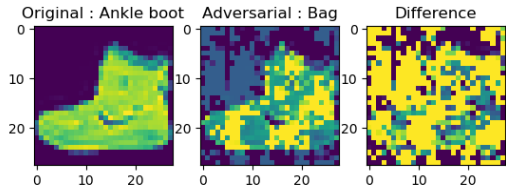
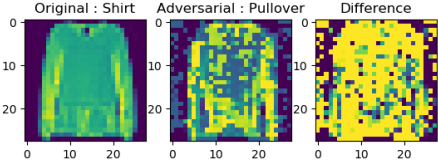
	L2 Distance: 58.638947
---	---------------------------

	49.684406
	53.834755
	54.681152
	52.770157

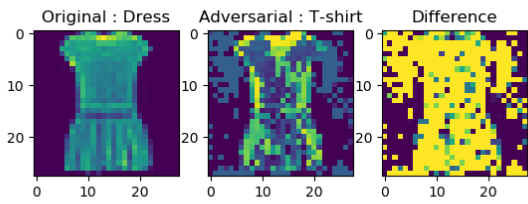
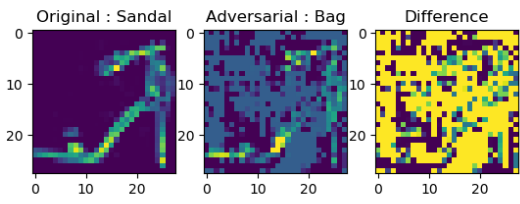
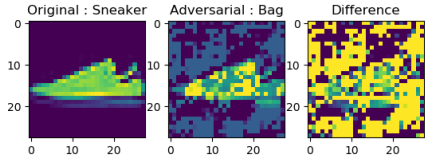
 <p>Three 28x28 heatmaps for digit 5. The 'Original' heatmap shows a clear yellow digit on a dark background. The 'Adversarial' heatmap shows the same digit with added noise and perturbations. The 'Difference' heatmap shows the difference between the original and adversarial images, with yellow and red areas indicating the perturbations.</p>	43.568077
 <p>Three 28x28 heatmaps for digit 6. The 'Original' heatmap shows a clear yellow digit on a dark background. The 'Adversarial' heatmap shows the same digit with added noise and perturbations. The 'Difference' heatmap shows the difference between the original and adversarial images, with yellow and red areas indicating the perturbations.</p>	56.02453
 <p>Three 28x28 heatmaps for digit 7. The 'Original' heatmap shows a clear yellow digit on a dark background. The 'Adversarial' heatmap shows the same digit with added noise and perturbations. The 'Difference' heatmap shows the difference between the original and adversarial images, with yellow and red areas indicating the perturbations.</p>	50.098133
 <p>Three 28x28 heatmaps for digit 8. The 'Original' heatmap shows a clear yellow digit on a dark background. The 'Adversarial' heatmap shows the same digit with added noise and perturbations. The 'Difference' heatmap shows the difference between the original and adversarial images, with yellow and red areas indicating the perturbations.</p>	53.93727

	<p>44.248154</p>
---	------------------

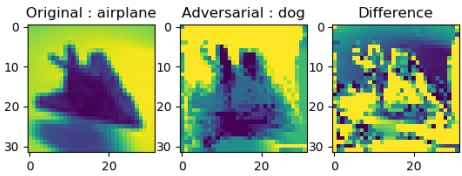
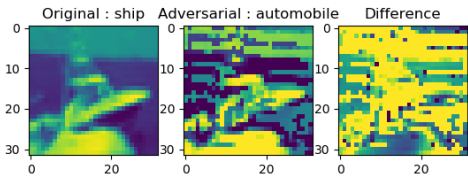
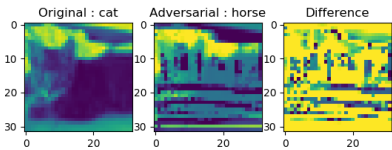
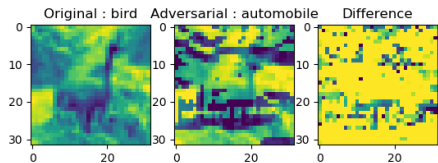
- Fashion MNIST

	<p>41.135635</p>
	<p>45.73863</p>
	<p>38.48382</p>

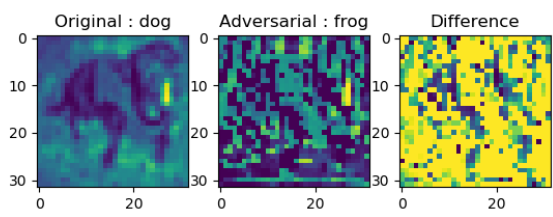
<p>Original : Pullover Adversarial : Shirt Difference</p>	41.70452
<p>Original : T-shirt Adversarial : Pullover Difference</p>	32.59449
<p>Original : Coat Adversarial : Bag Difference</p>	47.682644
<p>Original : Trouser Adversarial : Coat Difference</p>	38.61961

 <p>Original : Dress Adversarial : T-shirt Difference</p>	34.96115
 <p>Original : Sandal Adversarial : Bag Difference</p>	44.893368
 <p>Original : Sneaker Adversarial : Bag Difference</p>	46.37036

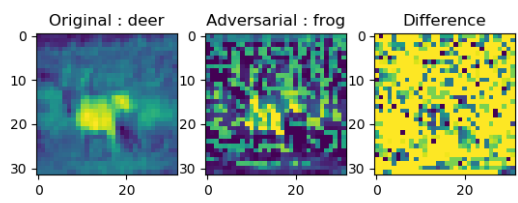
- Cifar10

	45.4934
	74.4539
	34.5965
	56.23459

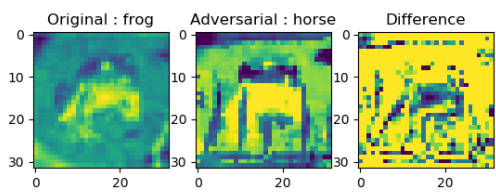
65.3459

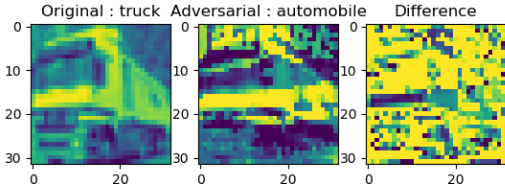
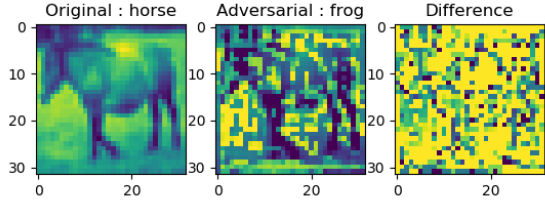
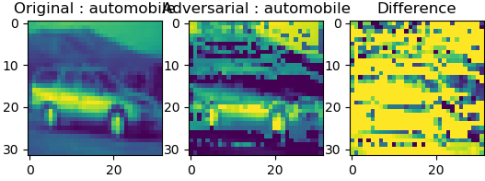


70.3495



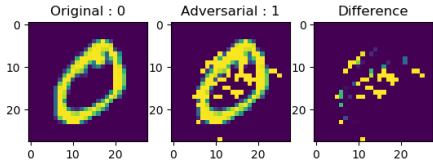
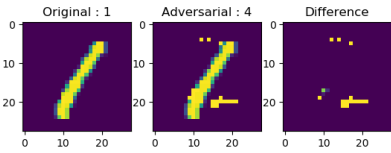
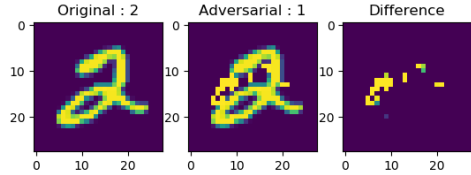
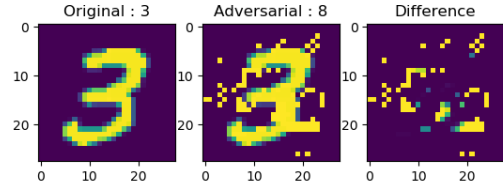
69.495

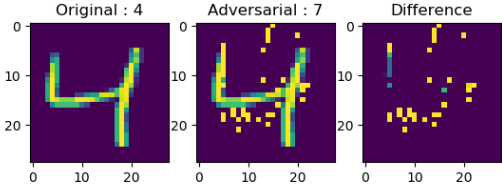
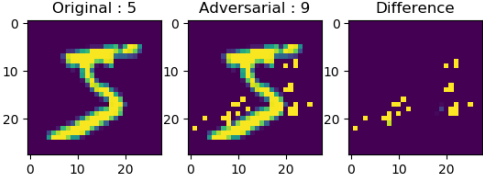
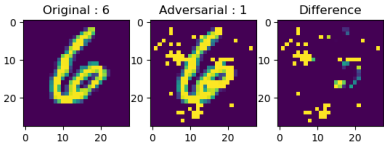
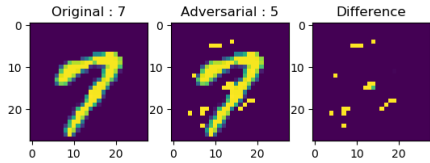


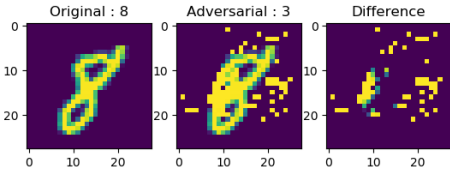
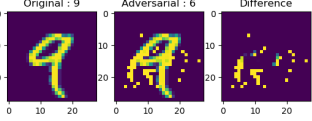
 <p>Original : truck Adversarial : automobile Difference</p>	55.546
 <p>Original : horse Adversarial : frog Difference</p>	34.2341
 <p>Original : automobile Adversarial : automobile Difference</p>	54.895

Saliency Map Attack

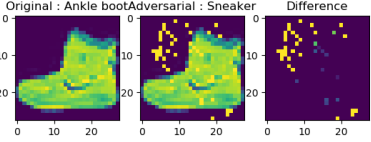
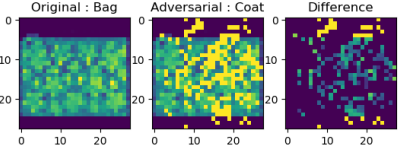
- MNIST

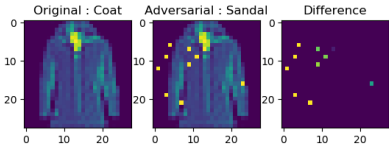
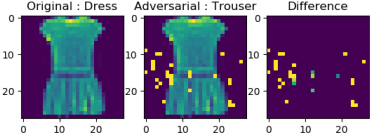
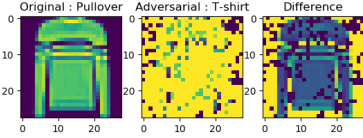
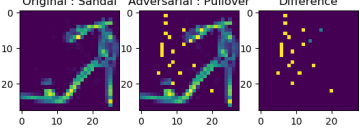
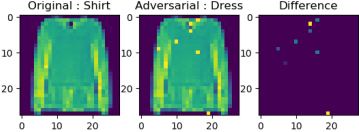
	<p>L2 Distance: 58.996033</p>
	<p>14</p>
	<p>13.958878</p>
	<p>28.186512</p>

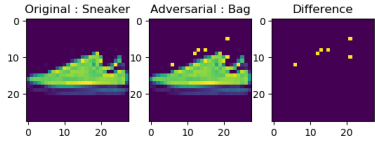
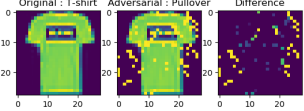
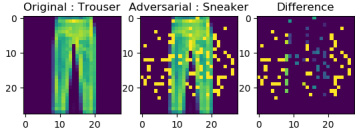
	22.35225
	40.974213
	19.889889
	17.790482

	68.6289
	10.442184

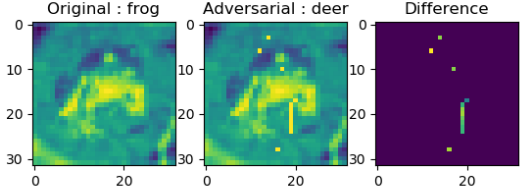
- Fashion Mnist

	12.42352
	32.3413

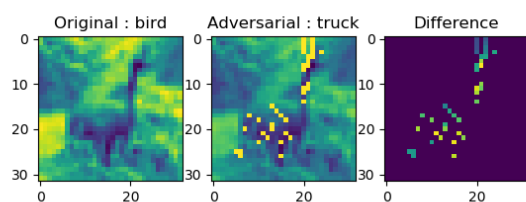
 <p>Original : Coat Adversarial : Sandal Difference</p> <p>The first heatmap shows a coat. The second heatmap shows the adversarial perturbation for the class 'Sandal'. The third heatmap shows the difference between the original and adversarial images.</p>	6.2145
 <p>Original : Dress Adversarial : Trouser Difference</p> <p>The first heatmap shows a dress. The second heatmap shows the adversarial perturbation for the class 'Trouser'. The third heatmap shows the difference between the original and adversarial images.</p>	10.1593
 <p>Original : Pullover Adversarial : T-shirt Difference</p> <p>The first heatmap shows a pullover. The second heatmap shows the adversarial perturbation for the class 'T-shirt'. The third heatmap shows the difference between the original and adversarial images.</p>	32.3413
 <p>Original : Sandal Adversarial : Pullover Difference</p> <p>The first heatmap shows a sandal. The second heatmap shows the adversarial perturbation for the class 'Pullover'. The third heatmap shows the difference between the original and adversarial images.</p>	12.65923
 <p>Original : Shirt Adversarial : Dress Difference</p> <p>The first heatmap shows a shirt. The second heatmap shows the adversarial perturbation for the class 'Dress'. The third heatmap shows the difference between the original and adversarial images.</p>	5.3896

 <p>Original : Sneaker Adversarial : Bag Difference</p>	6.34893
 <p>Original : T-shirt Adversarial : Pullover Difference</p>	13.1383
 <p>Original : Trouser Adversarial : Sneaker Difference</p>	18.3478

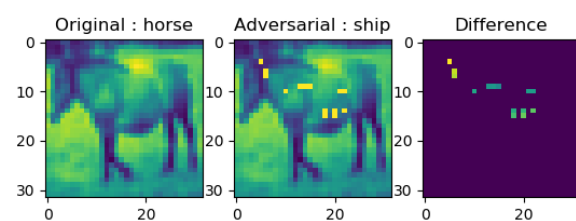
- Cifar10

 <p>Original : frog Adversarial : deer Difference</p>	L2 Distance: 3.293041
--	--------------------------

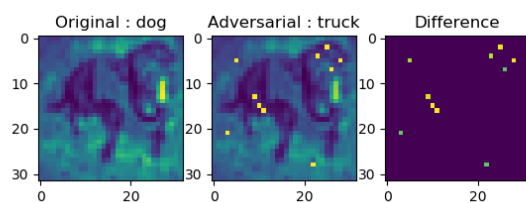
16.619839



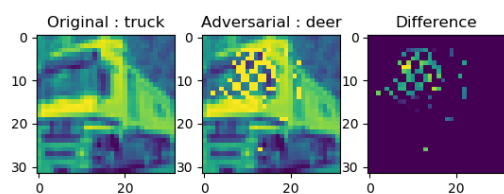
4.5920796



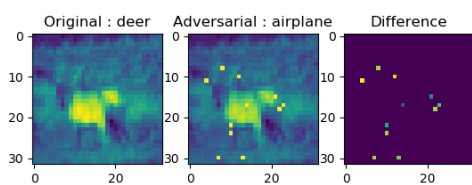
6.6491357



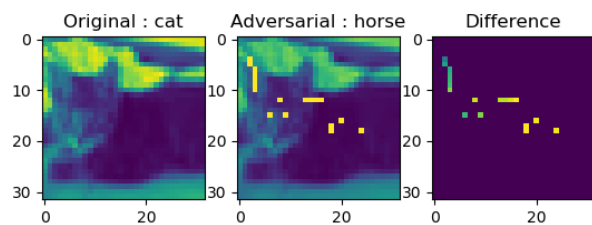
20.610796

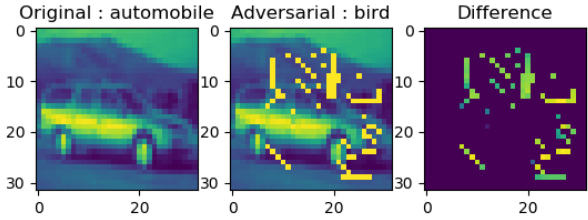
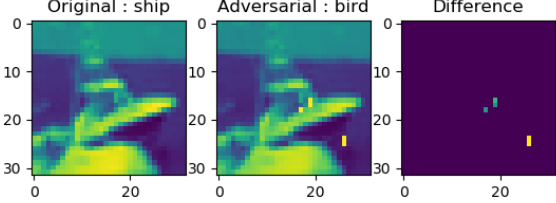
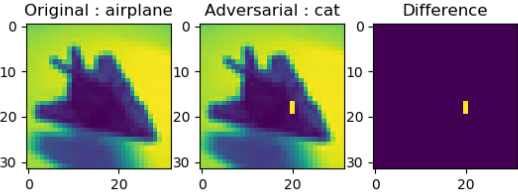


3.504764



10.67778

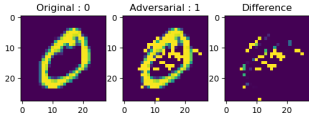
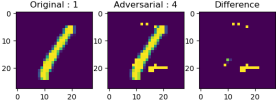
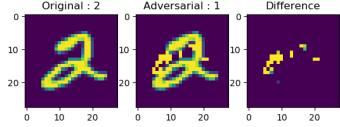
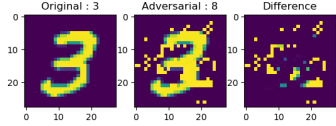
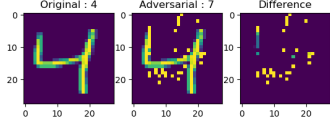
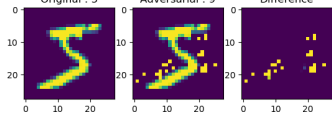


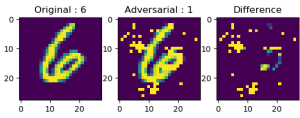
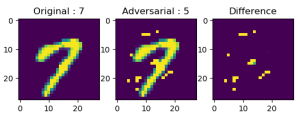
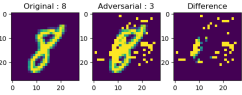
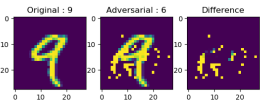
	49.69607
	2.267498
	2.710019

Saliency Map attack is able to introduce a small perturbation to the original image and effectively makes the model misclassified the objects. It has the smallest L2 distance between the original image and the adversarial image, compared to FGSM, MIM, and BIM.

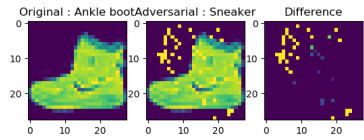
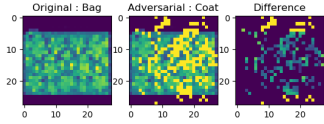
CW

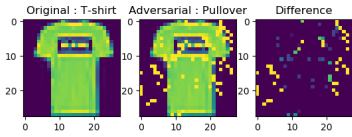
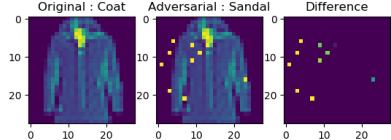
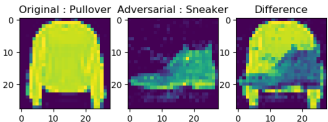
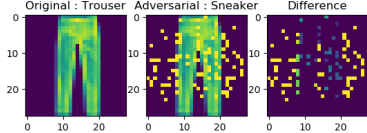
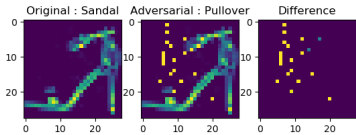
- MNIST

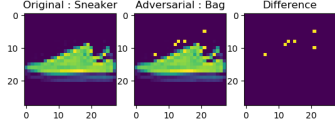
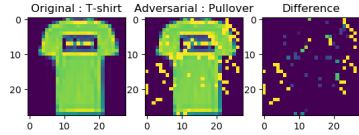
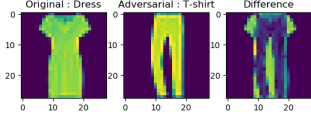
	9.5934
	5.3996
	6.348
	10.3996
	6.3492
	4.650

	9.6039
	5.069
	11.2954
	10.294

- Fashion MNIST

	15.395
	6.3493

 <p>Original : T-shirt Adversarial : Pullover Difference</p>	5.394
 <p>Original : Coat Adversarial : Sandal Difference</p>	5.92
 <p>Original : Pullover Adversarial : Sneaker Difference</p>	4.6923
 <p>Original : Trouser Adversarial : Sneaker Difference</p>	6.392
 <p>Original : Sandal Adversarial : Pullover Difference</p>	10.954

 <p>Original : Sneaker Adversarial : Bag Difference</p>	6.943
 <p>Original : T-shirt Adversarial : Pullover Difference</p>	6.1349
 <p>Original : Dress Adversarial : T-shirt Difference</p>	5.3920

From the figure above, the Carlini & Wagner L2 attack managed to make the classifier misclassified the object effectively by just introducing the smallest perturbation. The L2 distance between the adversarial image and the original image is relatively smaller compared to FGSM, BIM, MIM and SMA.

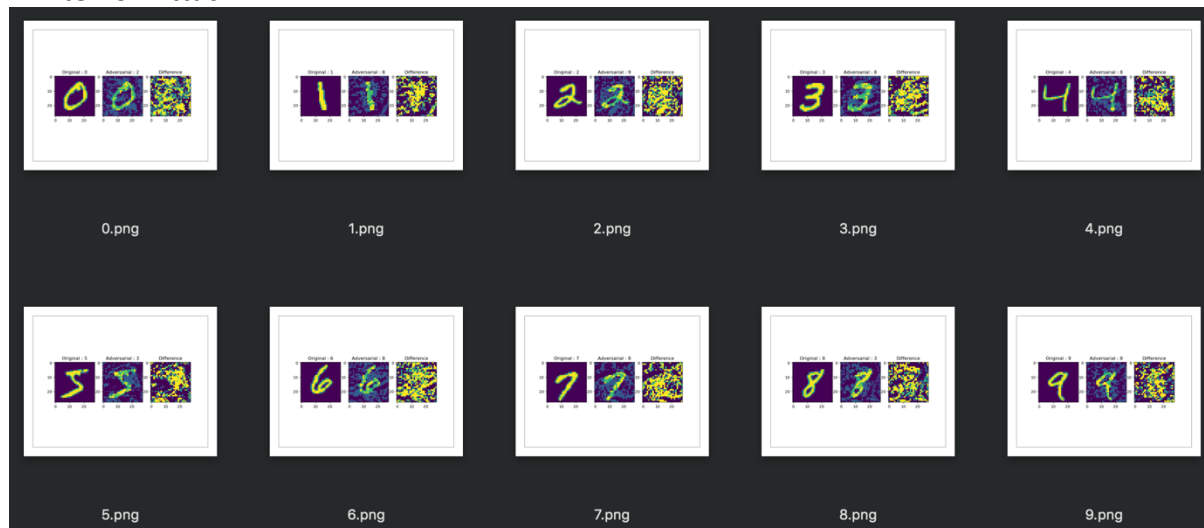
3) Analyze the result after applying a defense technique, namely adversarial training.

Dataset	Attack Method	Test accuracy on adversarial example	Test accuracy on adversarial example after defense techniques
MNIST	FGSM	0.1143	0.2254
MNIST	BIM	0.0059	0.1493
MNIST	MIM	0.0049	0.1175
FMNIST	FGSM	0.0631	0.1498
FMNIST	BIM	0.0636	0.0974
FMNIST	MIM	0.0643	0.1179
CIFAR10	FGSM	0.1295	0.6540
CIFAR10	BIM	0.0734	0.0994

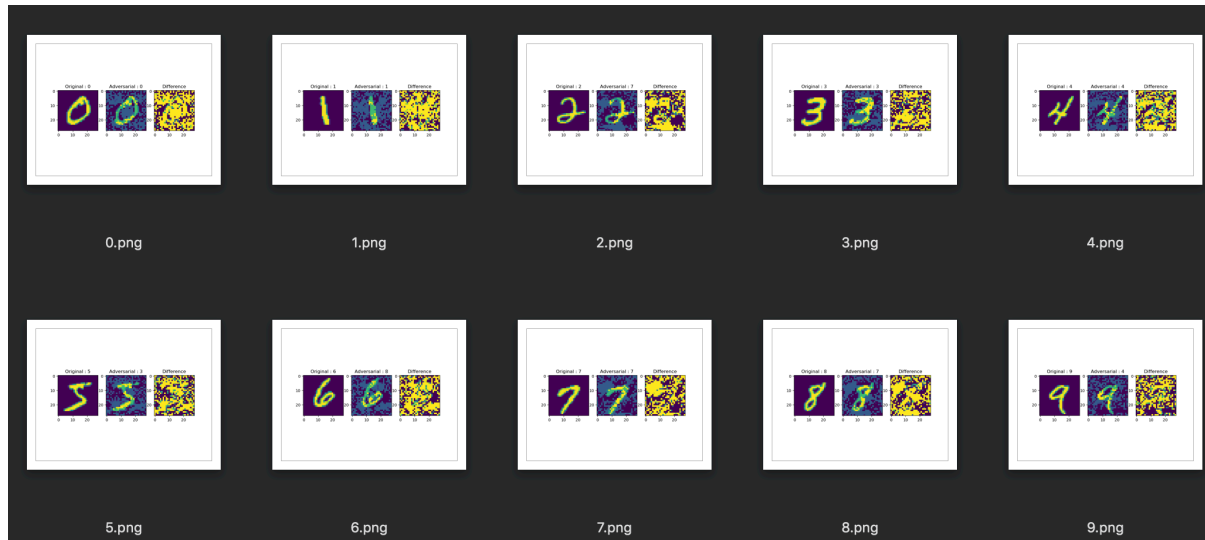
The application of defense against the adversarial examples is to make the model smoother by limiting the sensitivity to small perturbation of its inputs. Three types of attack (FGSM, BIM, MIM) are used against the MNIST, FMNIST, CIFAR10 datasets respectively. As the table illustrated, we can see that the accuracies increased after the application of the defense techniques for each test sample. The results show that the adversarial technique is able to increase the effectiveness of the adversarial model, especially the FGSM(CIFAR10), the accuracies was increased to 0.6540 from 0.1295.

4) Perform and analyze between a white-box attack and a black-box attack by using Basic Iterative Method attack technique.

White Box Attack:



Black Box Attack:



Black box trains a local model to substitute for the target DNN, using the inputs synthetically generated by an adversary and labeled by the target DNN. As black box suggested, it has no information about the structure or parameters of the DNN, and also does not have access to any large training dataset. It can only observe labels assigned by the designated model for chosen inputs. It segregates and trains a local substitute with a synthetic dataset – the inputs are synthetic and generated by the adversary while the outputs are the labels assigned by the target DNN and observed by the adversary. Black box strategy is to learn the substitute for the target model using a synthetic dataset generated by the adversary and labeled by observing the oracle output. Then, the adversarial examples are crafted using this substitute. We can see that the target DNN misclassified them due to the transferability between architecture.

For the white box attack, the intruder has the access to the model's parameters while the black box attacks the intruder has zero access to these parameters. It uses different model to generate adversarial images and run against the target model.

The results show that the white box attack using the basic iterative method achieves better results on the adversary, and also outperforms the ones with black box attack. The performance in black box attacks declined along with the number of epochs in the basic iterative method.

The dataset that I have targeted on was MNIST dataset, with the Basic Iterative Method. As shown in the figure above, the labeling between the input and the adversary does not make a significant mismatch on the accuracies in the black box attack. Whereas, in the white box attack, it has introduced adversarial noises which managed to fool the classifier to make the model misclassified label. In summary, the adversarial examples that generated by white box attack method are much tractable than the black box attack.