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.

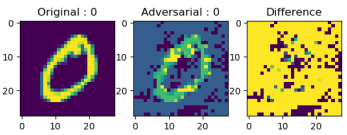
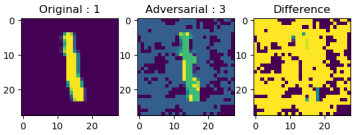
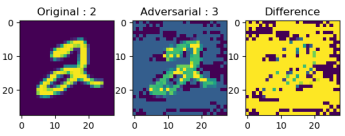
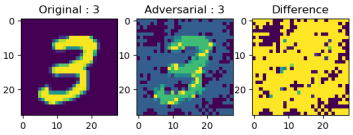
FGSM

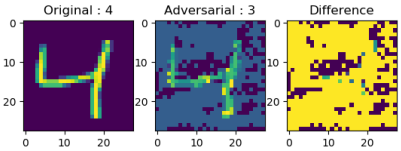
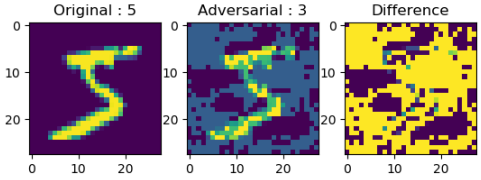
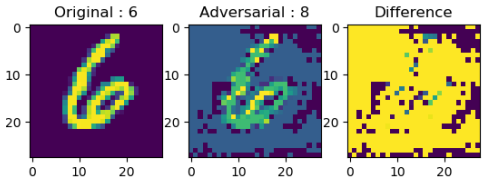
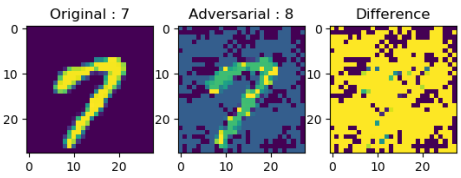
It is to linearize the cost function used to train a model around the neighborhood of the training point. The resulting adversarial example corresponding to the input x is computed as below:

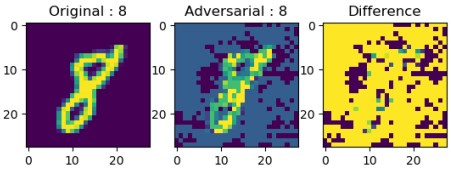
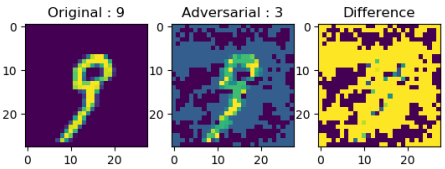
$$x \leftarrow x + \varepsilon \cdot \nabla J(f, \Theta, x)$$

- MNIST
  The L2 distance is calculated by

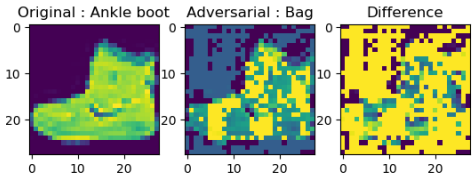
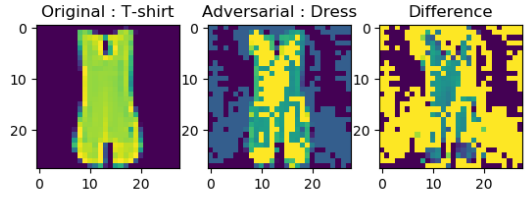
$$d(Adv, Ori) = \sqrt{[\Sigma(Adv[:,:] - Ori[:,:])]^2}$$

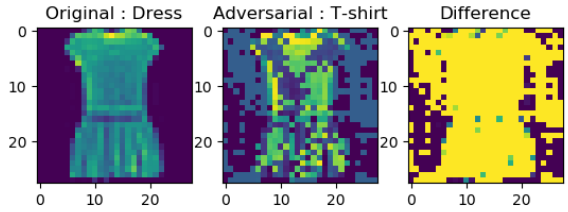
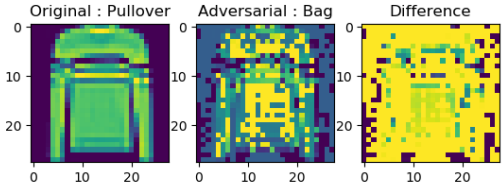
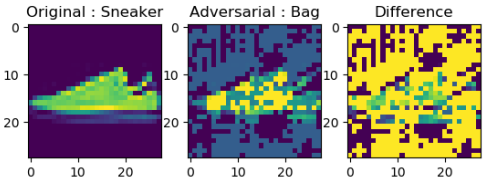
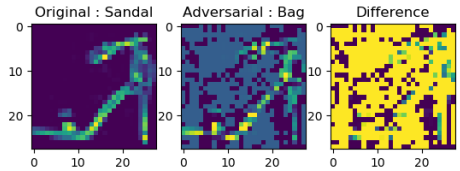
	L2 Distance: 58.638947
	49.684406
	53.834755
	54.681152

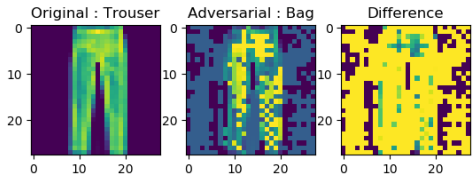
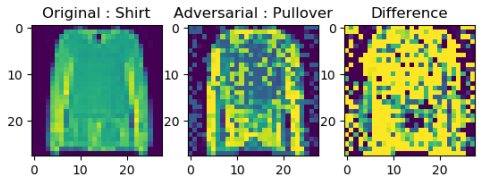
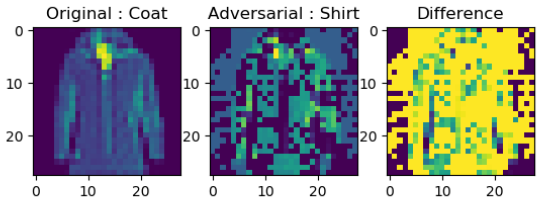
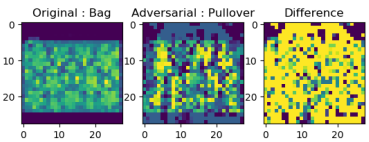
	52.770157
	43.568077
	56.02453
	50.098133

	53.93727
	44.248154

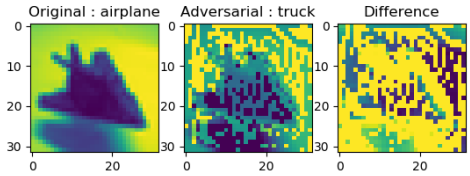
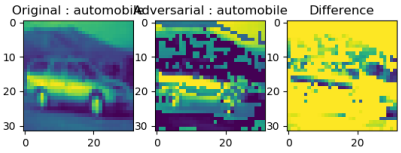
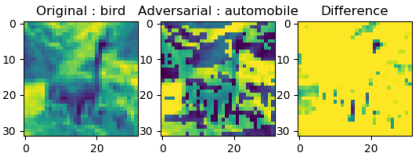
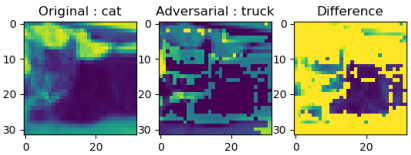
- Fashion MNIST

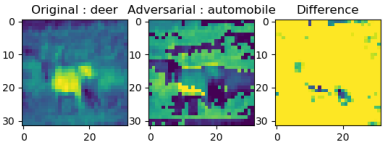
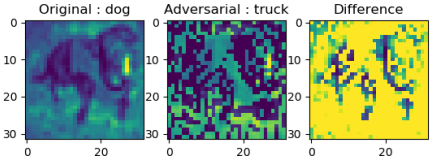
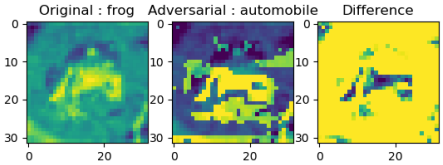
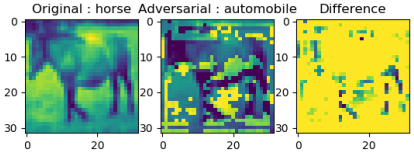
	L2 Distance: 49.731827
	52.33737

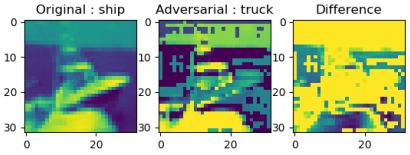
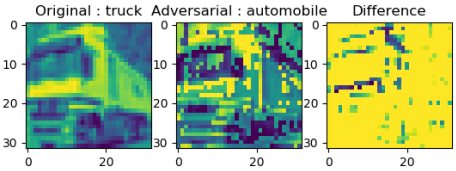
	52.33743
	57.61750
	46.305008
	49.135406

	52.868507
	58.084663
	50.613808
	58.124342

- Cifar10

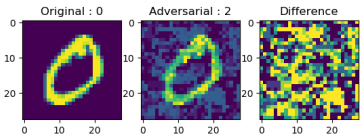
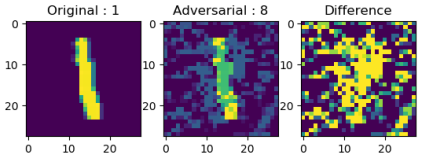
	L2 Distance: 58.34123
	60.342
	70.34234
	40.34237

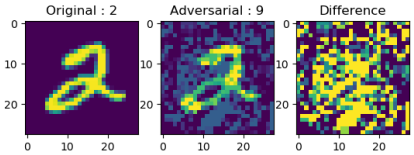
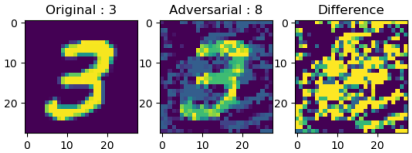
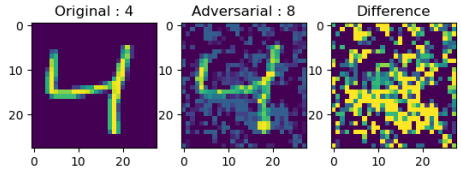
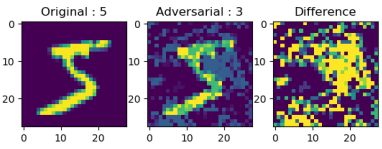
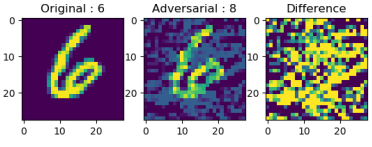
	70.59182
	60.23486
	65.45893
	58.65934

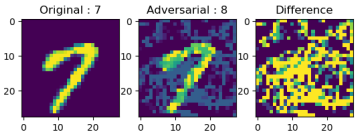
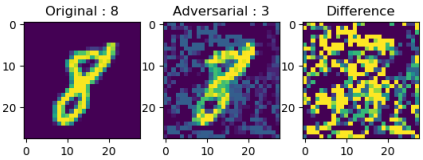
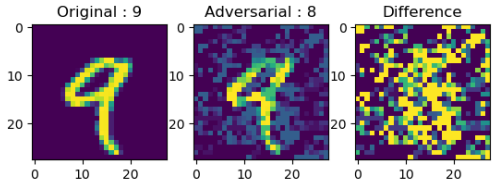
	50.83458
	58.485

Basic Iterative Method

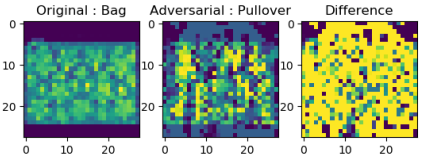
- MNIST

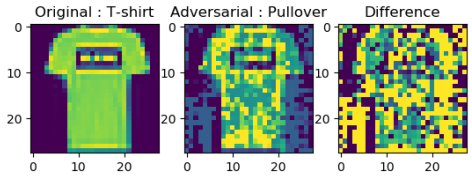
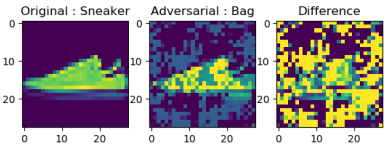
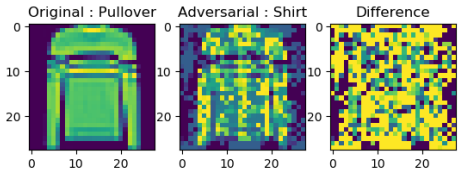
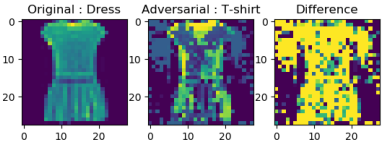
	L2 Distance: 68.2348
	69.3845

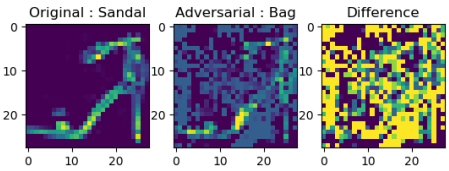
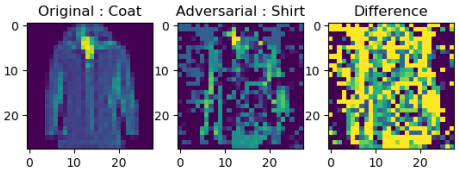
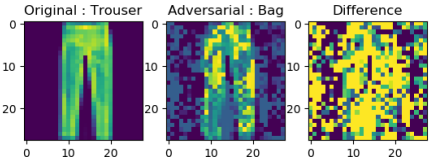
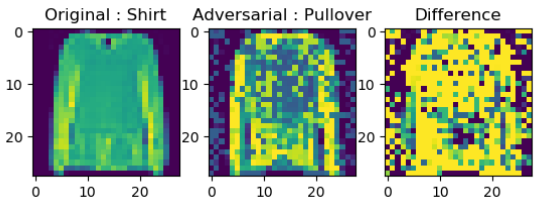
	65.2384
	20.459
	54.148
	44.349
	34.549

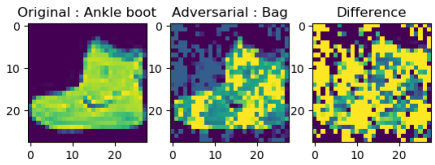
	49.4859
	45.48235
	33.12348

- Fashion MNIST

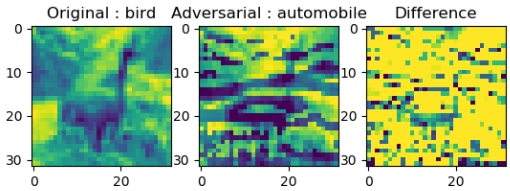
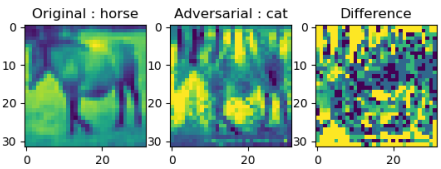
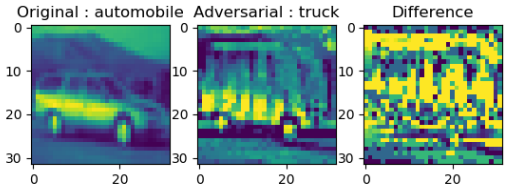
	L2 Distance: 36.150993

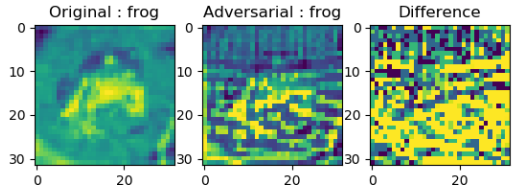
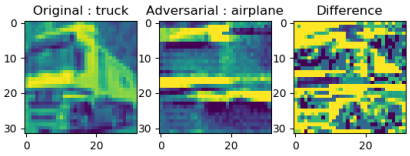
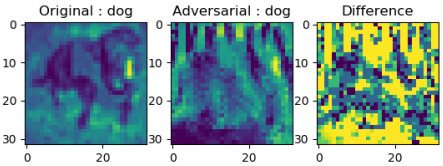
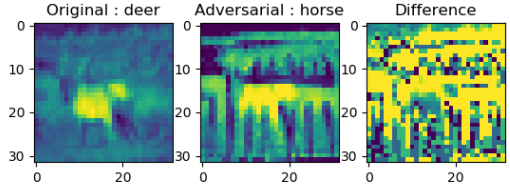
	35.113922
	32.874474
	
	38.022118

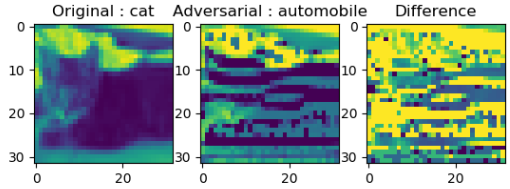
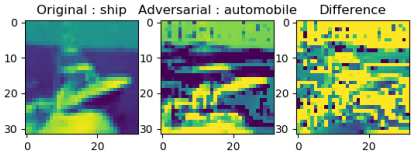
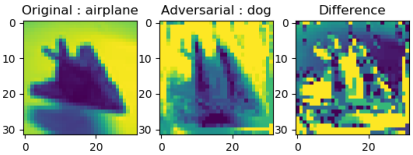
	29.258118
	40.22332
	32.052715
	32.017265

	33.219147

- Cifar10

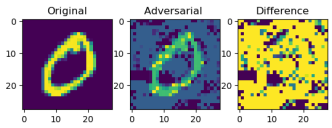
	L2 Distance: 67.3489
	65.34823
	79.3458

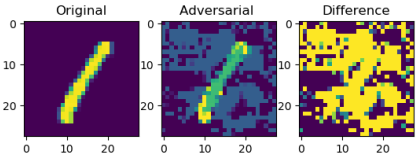
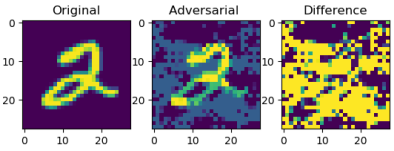
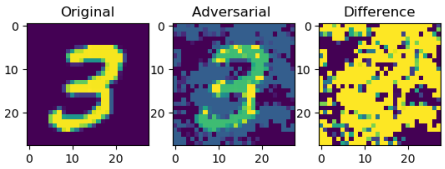
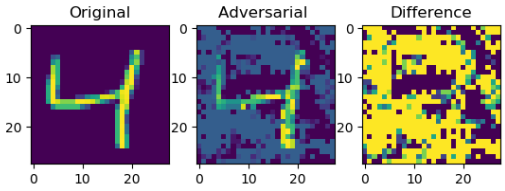
	68.34588
	78.3458
	68.3412
	54.2913

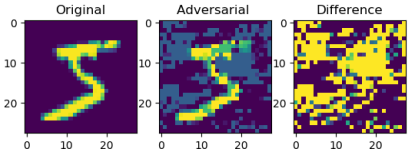
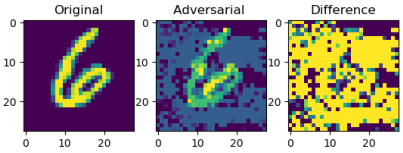
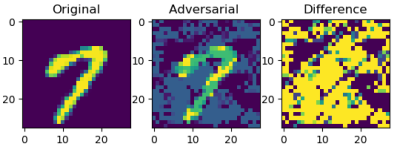
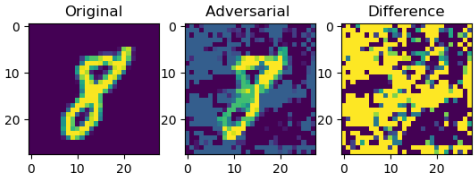
	32.495
	50.1923
	60.4903

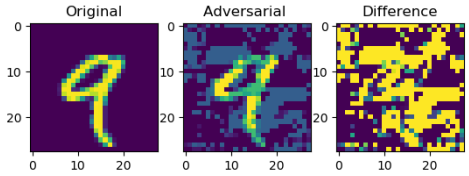
Momentum Iterative method

- MNIST

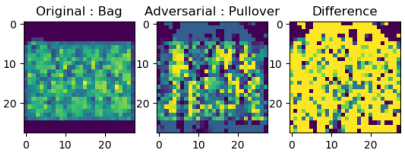
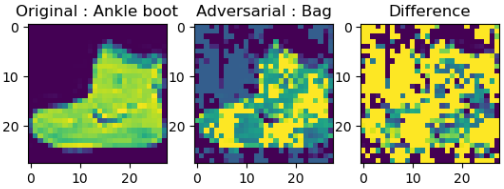
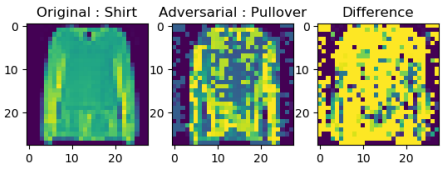
	L2 Distance: 58.638947

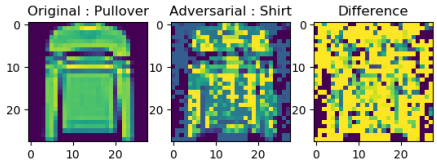
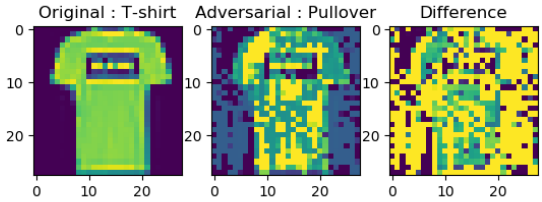
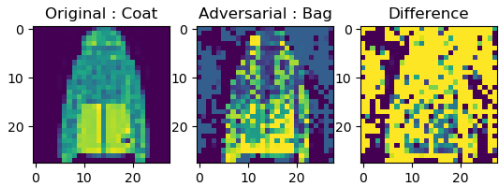
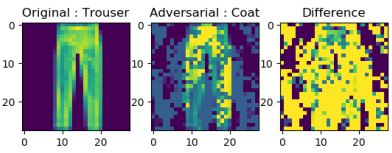
	49.684406
	53.834755
	54.681152
	52.770157

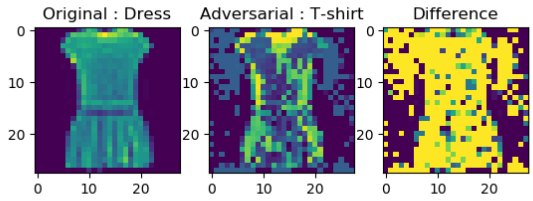
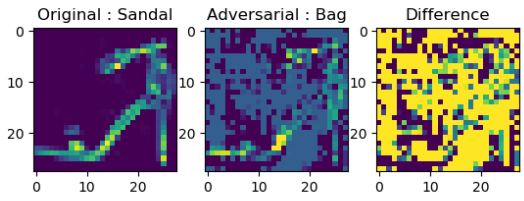
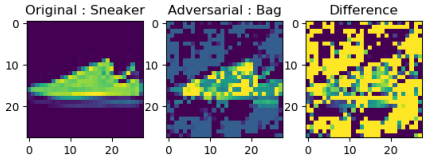
	43.568077
	56.02453
	50.098133
	53.93727

	44.248154

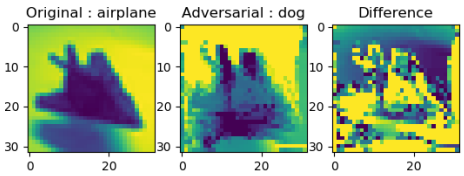
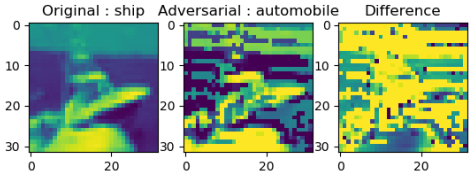
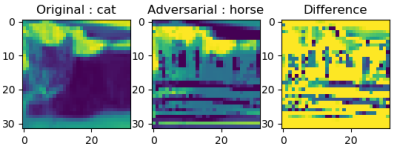
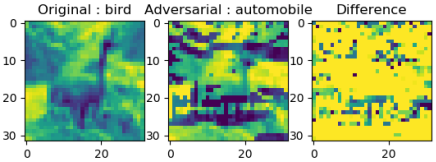
- Fashion MNIST

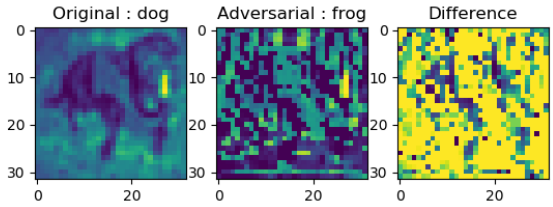
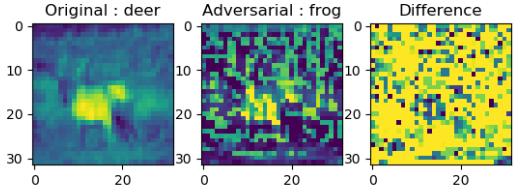
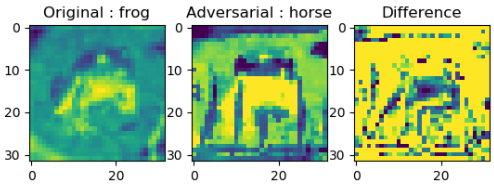
	41.135635
	45.73863
	38.48382

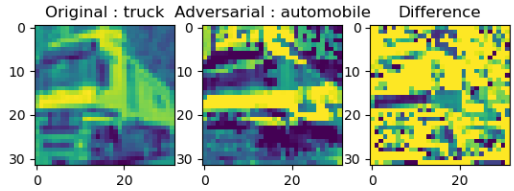
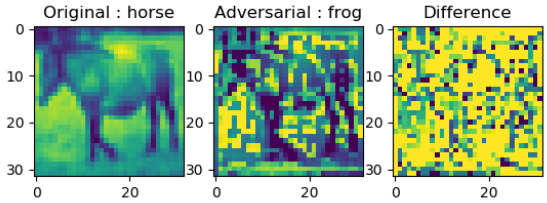
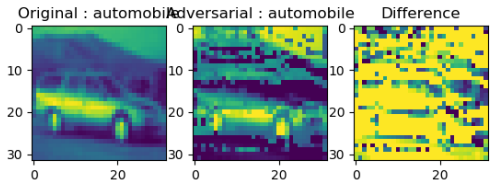
	41.70452
	32.59449
	47.682644
	38.61961

	34.96115
	44.893368
	46.37036

- Cifar10

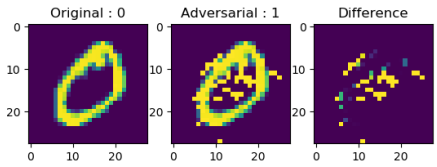
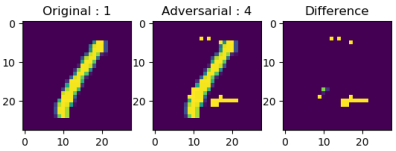
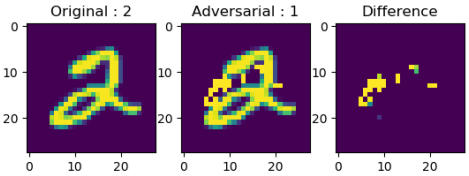
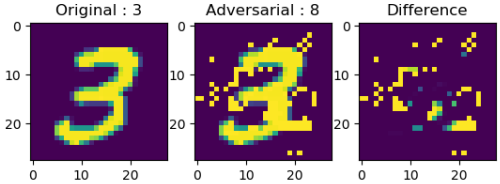
	45.4934
	74.4539
	34.5965
	56.23459

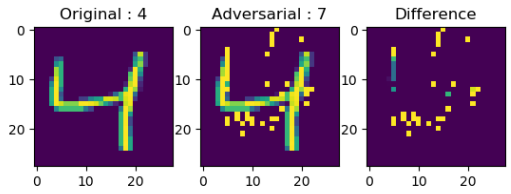
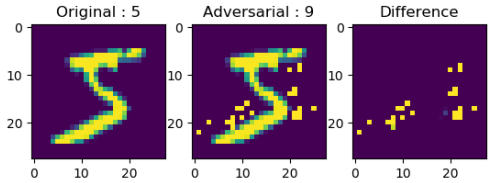
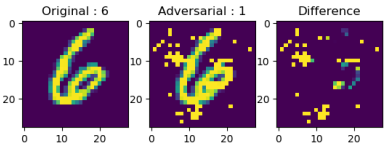
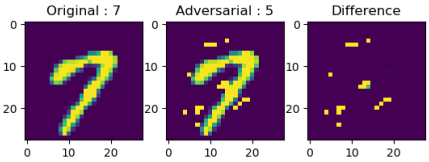
	65.3459
	70.3495
	69.495

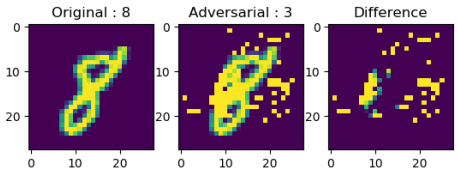
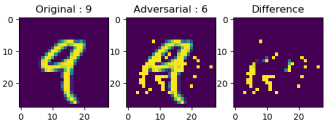
	55.546
	34.2341
	54.895

Saliency Map Attack

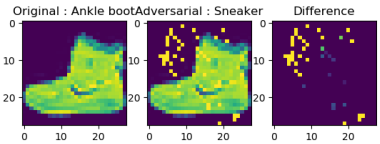
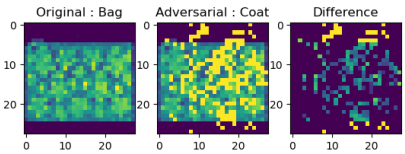
- MNIST

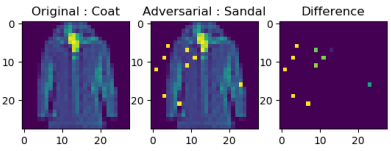
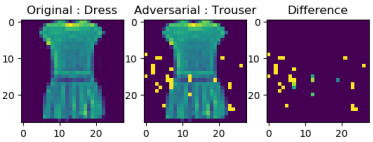
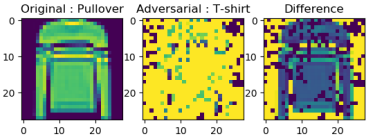
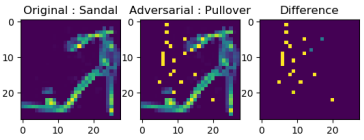
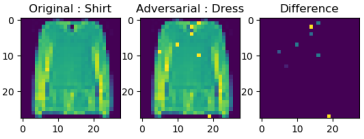
	L2 Distance: 58.996033
	14
	13.958878
	28.186512

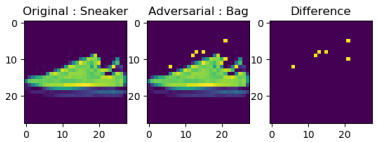
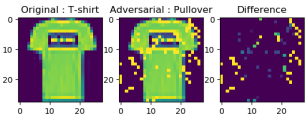
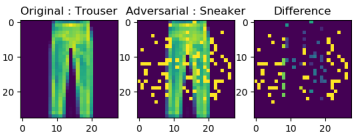
	22.35225
	40.974213
	19.889889
	17.790482

	68.6289
	10.442184

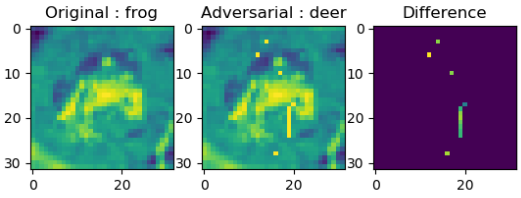
- **Fashion Mnist**

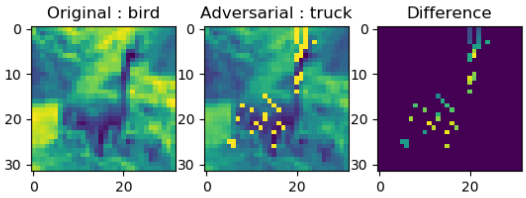
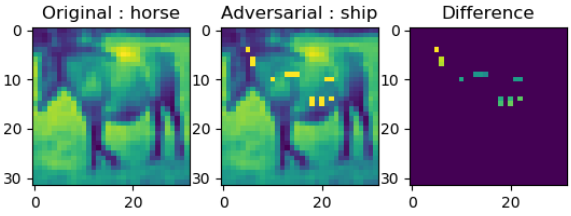
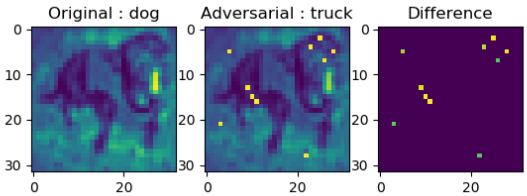
	12.42352
	32.3413

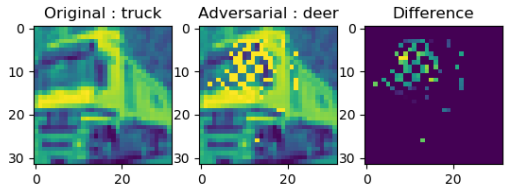
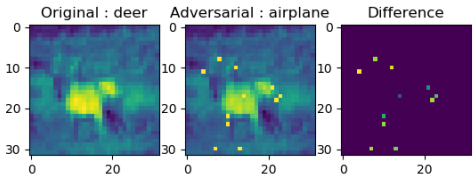
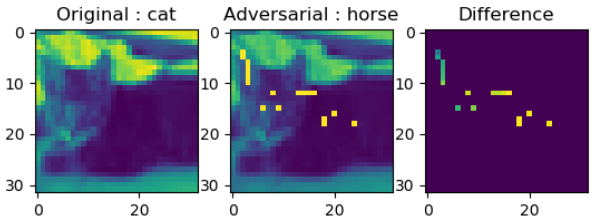
	6.2145
	10.1593
	32.3413
	12.65923
	5.3896

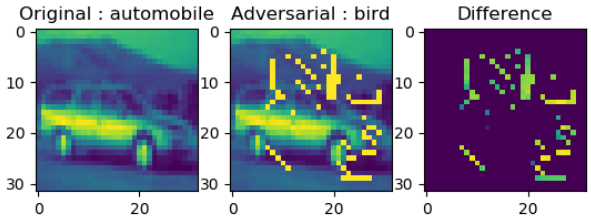
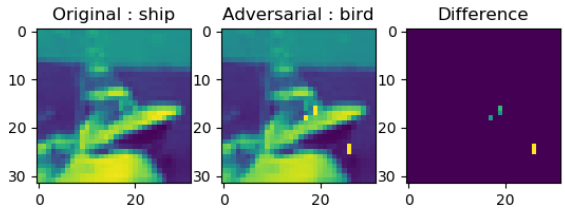
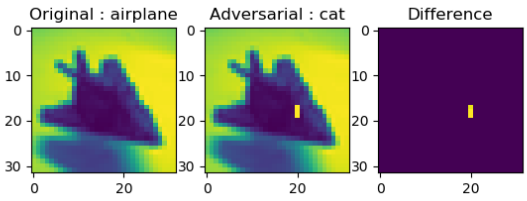
	6.34893
	13.1383
	18.3478

- **Cifar10**

	L2 Distance: 3.293041

	16.619839
	4.5920796
	6.6491357

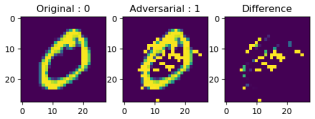
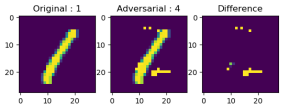
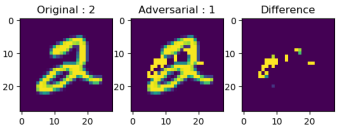
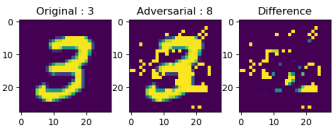
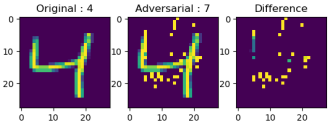
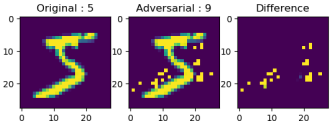
	20.610796
	3.504764
	10.67778

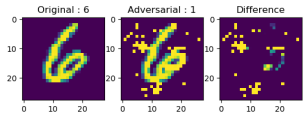
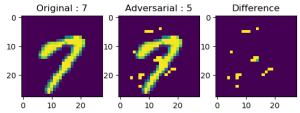
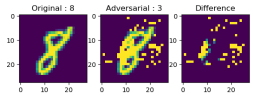
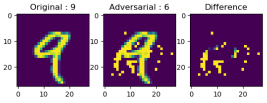
	49.69607
	2.267498
	2.710019

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.

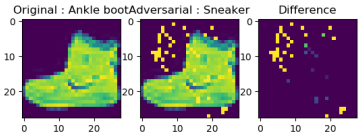
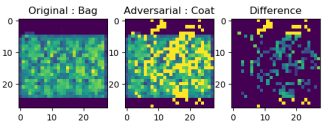
CW

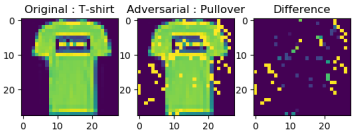
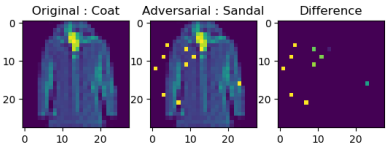
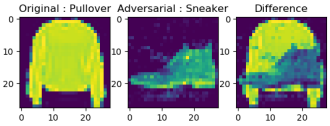
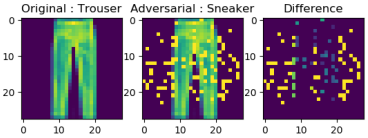
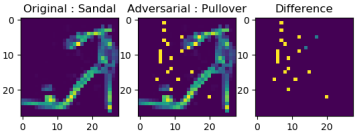
- MNIST

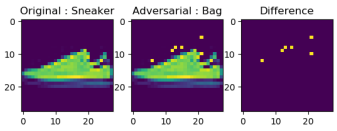
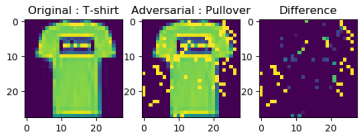
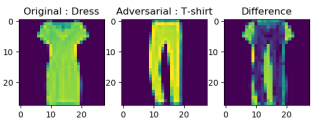
	9.5934
	5.3996
	6.348
	10.3996
	6.3492
	4.650

	9.6039
	5.069
	11.2954
	10.294

- Fashion MNIST

	15.395
	6.3493

	5.394
	5.92
	4.6923
	6.392
	10.954

	6.943
	6.1349
	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.

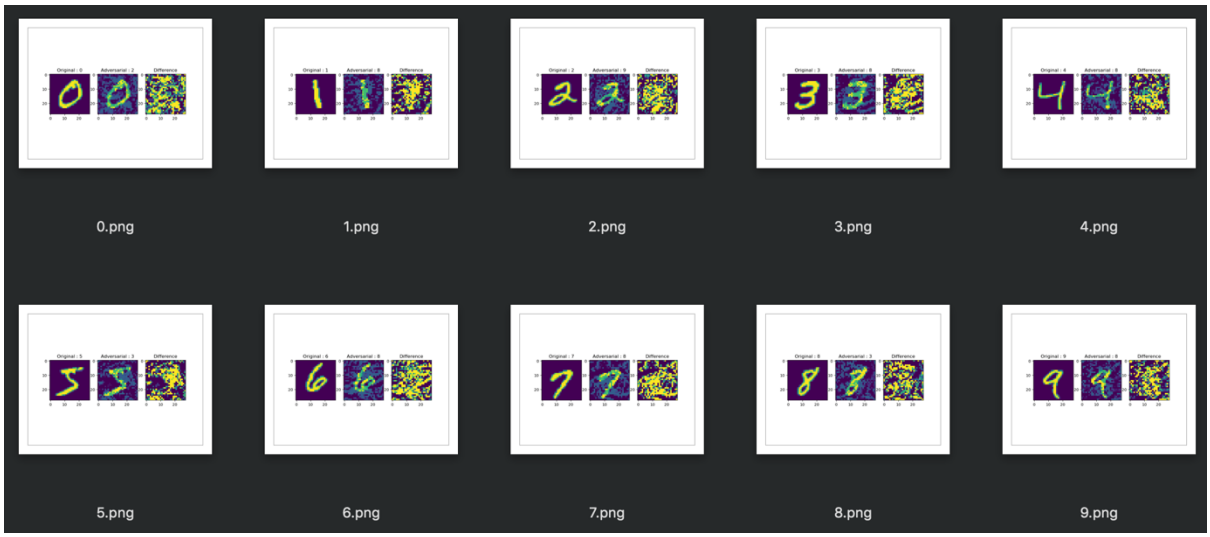
3) Analyze the result after applying a defense technique, namely adversarial training.

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.