CS579 – Homework 2 Solutions

Opeyemi Ajibuwa

Task 1

a. The classification accuracy of clean test-set samples on the 2 DNNs (LeNet and ResNet18) is presented in Table 1 below.

Test Sample Types	Model	Classification Accuracy
Clean	LeNet_MNIST	0.98872
	ResNet18_CIFAR10	0.87400
Adversarial	LeNet_MNIST	0.2499
	ResNet18_CIFAR10	0.1028

Table 1

Results analysis:

The results indicate that the accuracy of both models drops significantly when subjected to adversarial attacks. The accuracy of LeNet with MNIST drops from 98.87% for clean images to 24.99% for adversarial images. Similarly, the accuracy of ResNet18 with CIFAR10 drops from 87.40% for clean images to 10.28% for adversarial images. The results show that the PGD adversarial attack is effective at generating adversarial examples that are misclassified by LeNet and even the more complex ResNet18 architecture.

Task II

Sub-Task II.I

a. The plots of {epsilon} vs. {classification accuracy} on each of the 2 DNNs are shown in Figures 1a. The plotting data is summarized in Table 2 below.

Epsilon (fixed niter = 5)	0.01	0.02	0.03	0.04	0.05	0.1	0.2	0.3	0.4	0.5	1.0
Accuracy (LeNet)	0.3645	0.315	0.279	0.2499	0.2499	0.2499	0.2499	0.2499	0.2499	0.2499	0.2499
Accuracy (ResNet18)	0.1323	0.1119	0.1048	0.099	0.0974	0.0948	0.0911	0.0864	0.0832	0.0808	0.0674

Table 2

b. The plots of {# iterations} vs. {classification accuracy} on each of the 2 DNNs are shown in Figures 1b.

Iterations	1	2	3	4	5	10	20	30	40	80	100
Accuracy (LeNet)	0.3768	0.3341	0.3019	0.2755	0.2499	0.1516	0.0277	0.0028	0.0003	0.0	0.0
Accuracy (ResNet18)	0.1516	0.1333	0.1234	0.1111	0.1002	0.0839	0.0749	0.0704	0.07	0.0653	0.0662

Table 3

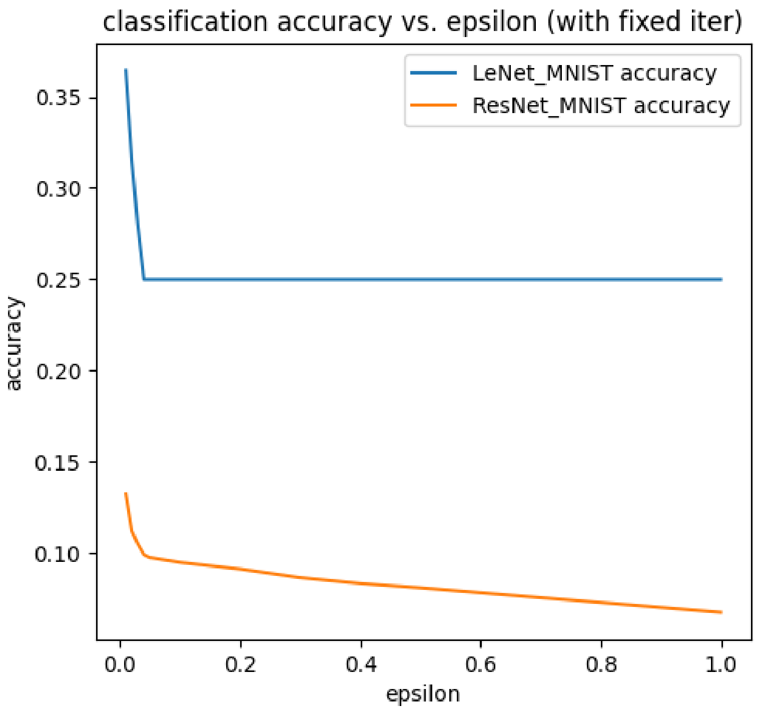


Fig. 1a: accuracy vs perturbation threshold

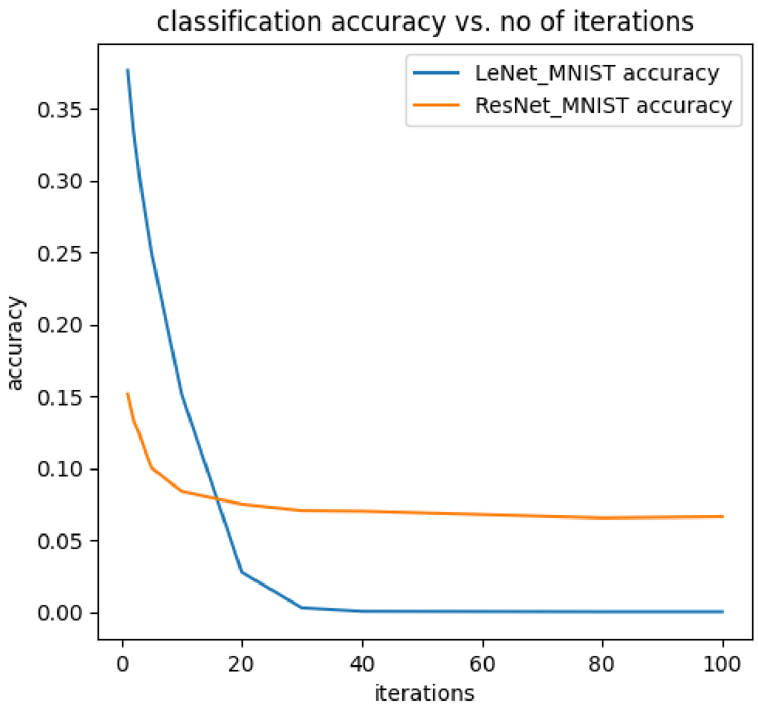


Fig. 1b: accuracy vs no of iterations

Results analysis:

The analysis of Figs. 1a demonstrates that, with a fixed iteration size and varying perturbation threshold ∈, the LeNet model performs better than ResNet18 under PGD attack. At smaller values of ∈, the LeNet model's accuracy decreases faster than ResNet's, and as the perturbation threshold increases, the LeNet model's accuracy remains fairly constant while ResNet's accuracy decreases linearly. This difference in behavior between the two models may be explained by the fact that the CIFAR10 dataset used with the ResNet model is more complex than the simpler MNIST dataset used with the LeNet model.

In Fig. 1b, it can be observed that the LeNet model initially achieves higher classification accuracy than ResNet. However, as the number of iterations increases, the accuracy of the LeNet model drops sharply and rapidly approaches zero within 30 iterations. This suggests that more iterations do not necessarily degrade the classification performance of the LeNet model. On the other hand, the accuracy of the ResNet model drops over 20 iterations and then remains approximately constant up to the 100th iteration. This indicates that the ResNet model's accuracy is quickly degraded over short iteration sizes, beyond which its performance remains fairly constant.

Sub-Task II.II

a. Using the same default hyperparameters as in task 1, the results of crafting adversarial examples on the test-samples for the two models with and without data augmentations are summarized in Table 3 below.

Data augmentations	**Model**	**Classification Accuracy**
With random rotations	LeNet_MNIST	0.2281
	ResNet18_CIFAR10	0.0728
With horizontal flips	LeNet_MNIST	0.2058
	ResNet18_CIFAR10	0.1038
With random rotations and horizontal flips	LeNet_MNIST	0.1917
	ResNet18_CIFAR10	0.0723
Without random rotations **OR** Without horizontal flips **OR** Without random rotations and horizontal flips	LeNet_MNIST	0.2499
	ResNet18_CIFAR10	0.1029

Table 4

b. With/without dropout regularization on the ResNet18-CIFAR10 model tested on adversarial samples, we have the following results summarized in Table 5.

Model	**Classification Accuracy**
ResNet18_with_dropout	0.0349
ResNet18_without_dropout	0.1028

Table 5

c. The results of the classification accuracy with the different weight decays are summarized in the table below.

	Classification accuracy on clean samples	**Classification accuracy on adversarial samples**
model with weight decay = 1e-5	0.30691	0.0271
model with weight decay = 1e-4	0.10164	0.0715
model with weight decay = 1e-3	0.10304	0.0566
model with weight decay = 1e-2	0.09964	0.1
model with weight decay = 1e-1	0.09994	0.1

Table 6

Results analysis:

Table 4 demonstrates that both the LeNet and ResNet models had reduced classification accuracy when subjected to random rotations. While horizontal flips also decreased the accuracy of the LeNet model, the ResNet model was not affected by this transformation. When random rotations and horizontal flipping were combined, the accuracy of both models slightly decreased. Although these transformations had varying effects on the classification accuracy of the two models, none of them was able to mitigate the adversarial attacks against the models.

Meanwhile, Table 5 shows that the ResNet model with dropout did not improve the classification accuracy of the model over the adversarial examples. In fact, the accuracy of the ResNet model with dropout worsened compared to the model without dropout. This observation contrasts with the expectation that ResNet with dropout reduces overfitting caused by the PGD attack.

For the weight decay regularization, results in Table 6 shows that as weight decay is decreased, the ResNet model's classification accuracy increases for clean samples. The opposite is the case on the adversarial samples where classification accuracy increases with increasing weight decay. This observation suggests that weight decay regularization may help prevent overfitting and improve the generalization performance of ResNet on adversarial examples, although the effect is relatively small.

Task III

a. Table 7 below summarizes the results of the models' accuracy on adversarial examples compared to the undefended models.

Model	**Classification Accuracy on adversarial examples**	
	Attack iterations = 5	**Attack iterations = 7**
LeNet_MNIST_undefended	0.2499	0.211
LeNet_MNIST_adv	0.8406	0.6849

Model	Classification Accuracy on adversarial examples	
	Attack iterations = 5	Attack iterations = 7
ResNet_CIFAR10_undefended	0.1028	0.0629
ResNet_CIFAR10_adv	0.1992	0.1586

Table 7

b. Table 8 below summarizes the results of the models' accuracy on clean test samples compared to the undefended models.

Model	Classification Accuracy on clean examples	
	Attack iterations = 5	Attack iterations = 7
LeNet_MNIST_undefended	0.98872	
LeNet_MNIST_adv	0.96915	0.98243
ResNet_CIFAR10_undefended	0.87400	
ResNet_CIFAR10_adv	0.32368	0.16623

Table 8

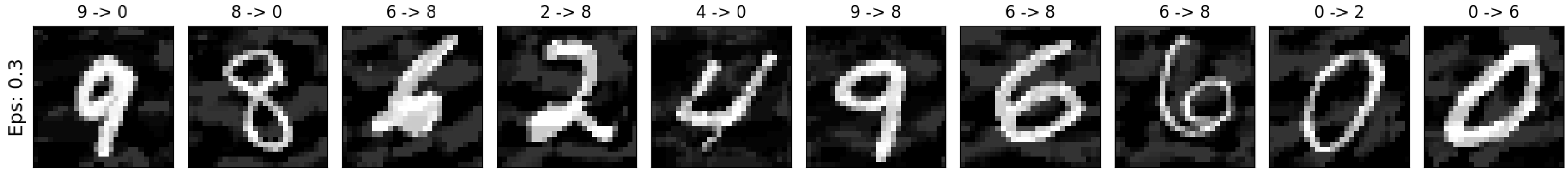
Results analysis:

Table 7 shows that increasing the number of attack iterations from 5 to 7 results in reduced classification accuracy for both undefended LeNet and ResNet models on adversarial examples. Similarly, the accuracy of the adversarially trained models also decreased with more iterations. This aligns with our expectation that classification performance deteriorates with increasing iterations. However, the results in Table 7 clearly indicate that the adversarially trained models outperform the undefended models on the adversarial examples.

In contrast, as Table 8 shows, the classification accuracy of undefended models on clean samples remained unchanged irrespective of the number of attack iterations. However, increasing the attack iterations from 5 to 7 improved the classification accuracy of the adversarially trained LeNet model on clean samples, while it decreased the accuracy of the adversarial ResNet model. The decrease in the classification accuracy of the ResNet model is consistent with the prior observations.

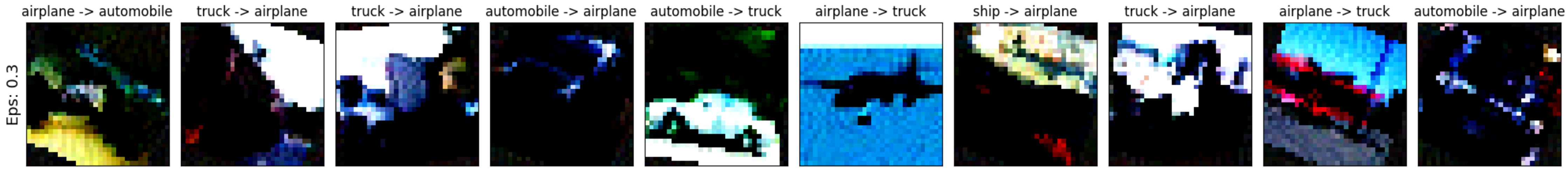
Bonus Task:

For MNIST: My trained DNN classification results are shown at the top of the images while Google Vision API best predictions are shown beneath the images.



9 → 9	8 → 8	8 → head	2 → 2	2 → sport equipment	9 → b	6 → 6	6 → 6	0 → o	0 → 0

For CIFAR10: My trained DNN classification results are shown at the top of the images while Google Vision API best predictions are shown beneath the images.



Packaged goods	Ronin	Watch	2D barcode	2D barcode	2D barcode	Rectangle	Rectangle	2D barcode	2D barcode