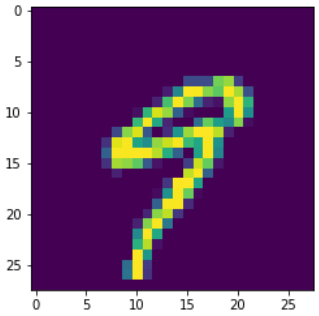
CIS 680: Project 1 Part B

Junfan Pan

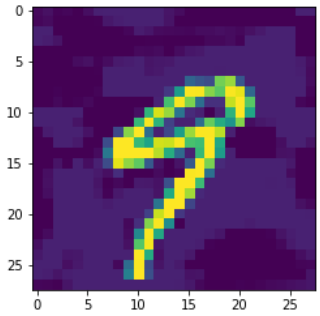
09/19/2020

1 Adversarial Images

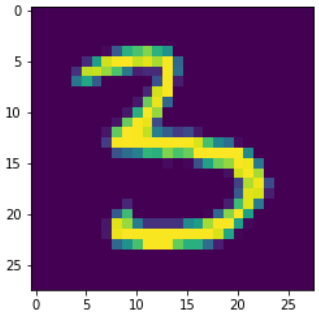
1.1 Untargeted Attack with pre-trained neural network



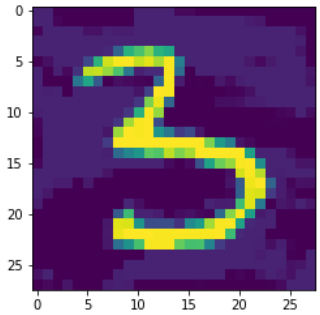
(a) original image



(b) untargeted attack image



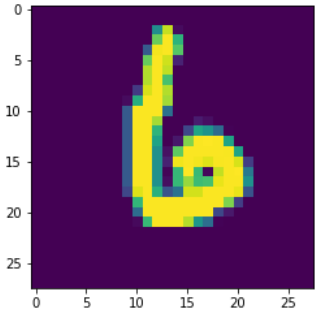
(a) original image



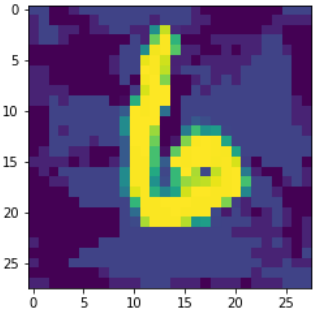
(b) untargeted attack image

- Describe the difference between adversarial images and original images
  In the adversarial images, some pixels around the digits become lighter which let the network to misclassify as different digits.

1.2 Targeted Attack with pre-trained neural network



(a) original image



(b) targeted attack image with target value as 4

- In the adversarial images with targeted attack, more pixels around the digits in the whole image become lighter comparing with those adversarial images generated by untargeted attack.

1.3 Targeted and Untargeted Attack using pre-trained neural network and re-trained neural network

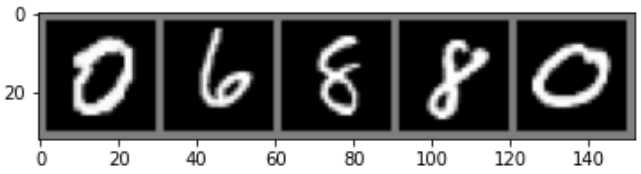
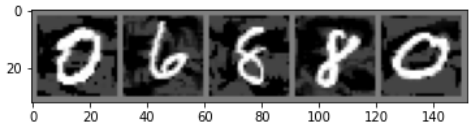
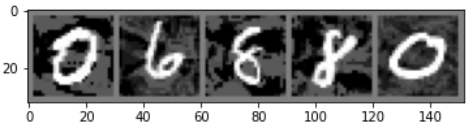


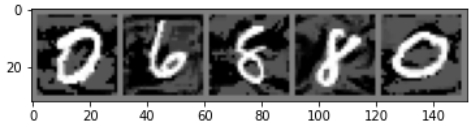
Figure 4: original image with all correct prediction as 0,6,8,8,0



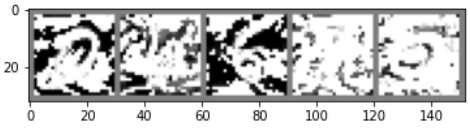
(a) untargeted attack image using pre-trained network with prediction as 2, 2, 6, 4, 2



(b) targeted attack image using pre-trained network with target value 2, 4, 6, 4, 4



(a) untargeted attack image using re-trained network with prediction as 2, 3, 6, 6, 9



(b) targeted attack image using re-trained network with target value as 2, 4, 6, 4, 4

- Comparing the adversarial images generated by untargeted attack and targeted attack, we can see that it is easier to achieve the former one.

- Networks trained using the same structures with same training datasets but with different initializations, the adversarial images generated from the same test examples can still be different. In addition, if we want to use the targeted attack with the same target value, it is possible that the previous network can easily achieve it while for the re-trained network it can be very difficult. We can observe that the adversarial images generated using targeted attack with re-trained network are almost all blurred and we almost can't see anything.