Adversarial Learning

Home Assignment III

Ido Calman, Matan Levy

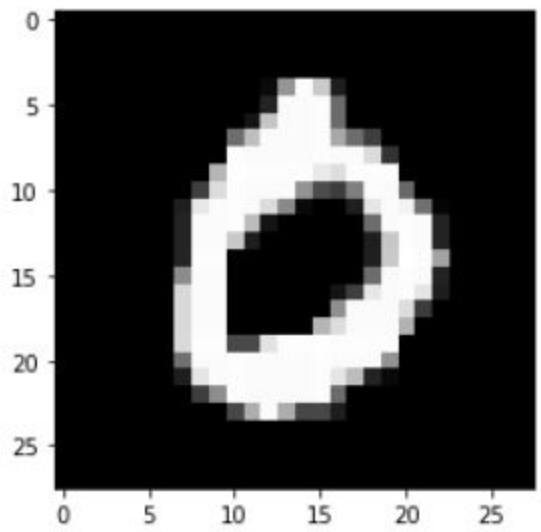
3. Non-defended results:

Attack	Success	Mean Perturbation Distance
FGSM	0.93	0.23
TGSM	0.47 (move out: 0.79)	0.23
PGD Untargeted	0.95	0.14
PGD Targeted	0.72 (move out: 0.75)	0.14

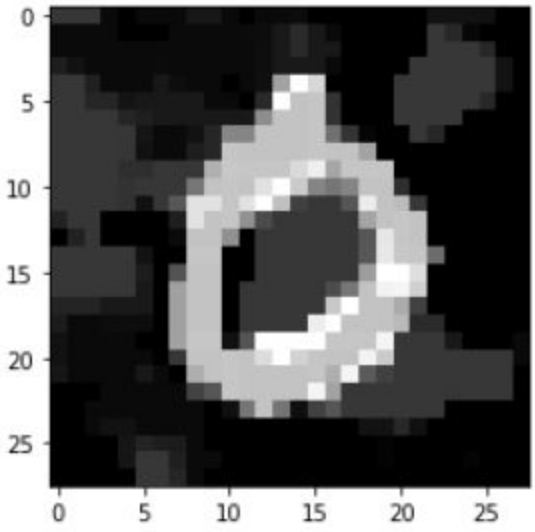
5. Defensive Distillation results:

Attack	Success Rate	Mean Perturbation Distance
FGSM	0.26	0.08
TGSM	0.5 (move out: 0.91)	0.29
PGD Untargeted	0.27	0.04
PGD Targeted	0.62 (move out: 0.69)	0.14

PGD untargeted example:



PGD targeted example:



Clearly, since gradients are being zeroed out in Defensive Distillation, PGD is having a hard time perturbing the image altogether. This phenomenon is clearer on the untargeted example, where almost no pixels are being perturbed from the original image. Similarly to the discussion in class, the Defensive Distillation approach does not have such a large effect on targeted attacks, so the degradation in success rate is as expected.

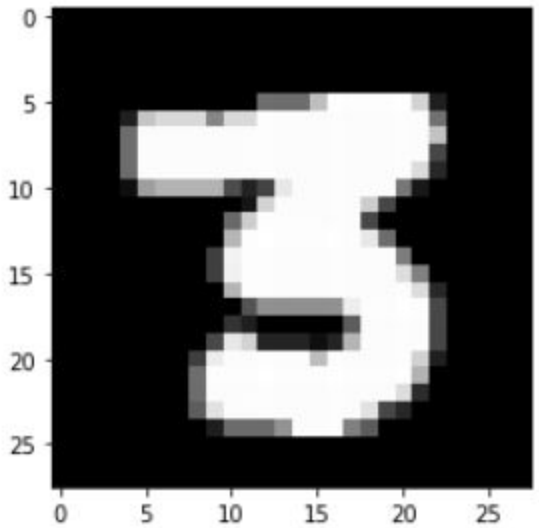
6. Attacking the binary distilled model:

Attack	Success Rate	Mean Perturbation Distance
FGSM	0.08	0.02
TGSM	**0.88**	0.29
PGD Untargeted	0.08	0.01
PGD Targeted	0.48	0.14

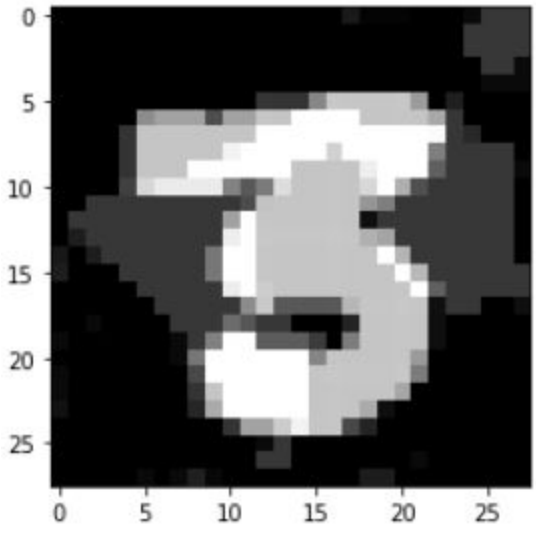
As discussed in class, the distilled defensive method practically works for untargeted attacks and struggle against targeted attacks.

Note: move out success rate equals exactly to the success rate of a targeted attack since moving out from one binary class to another is the same as succeeding in targeting the opposite class.

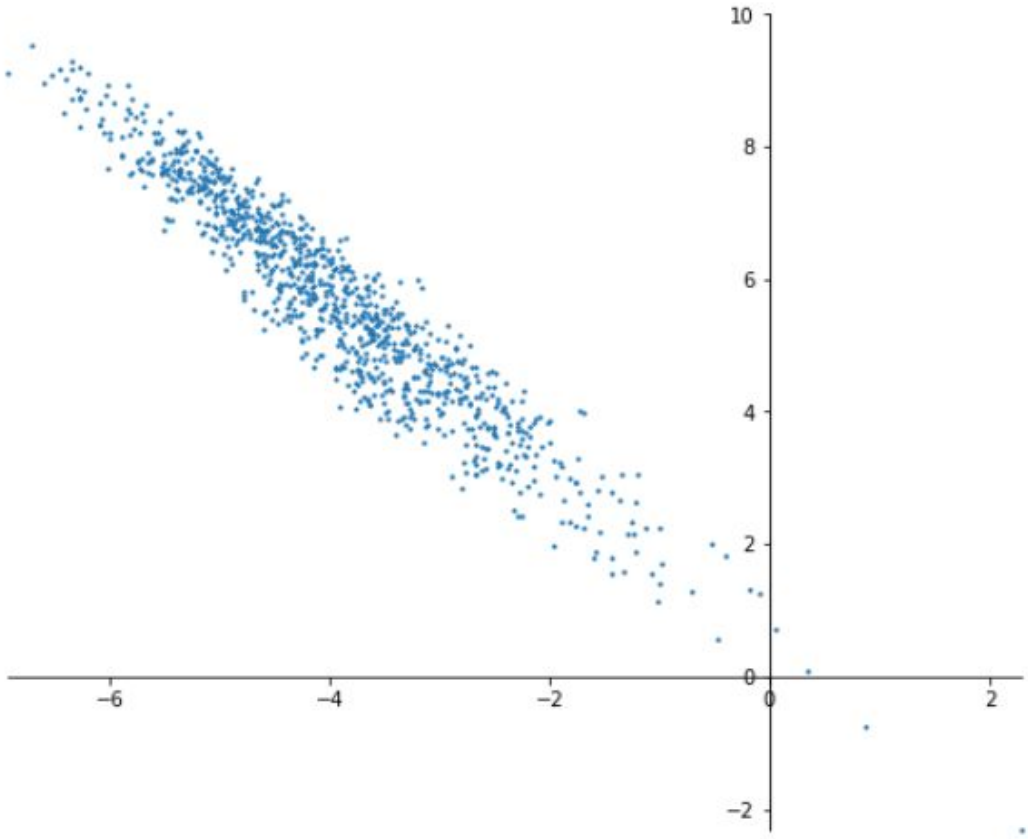
Untargeted example



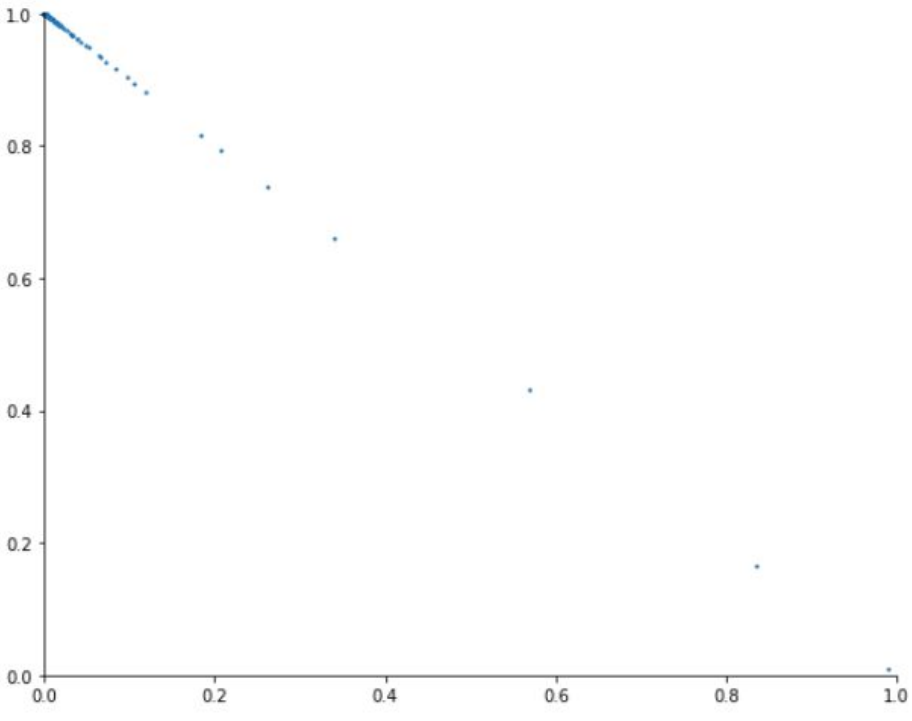
Targeted Example



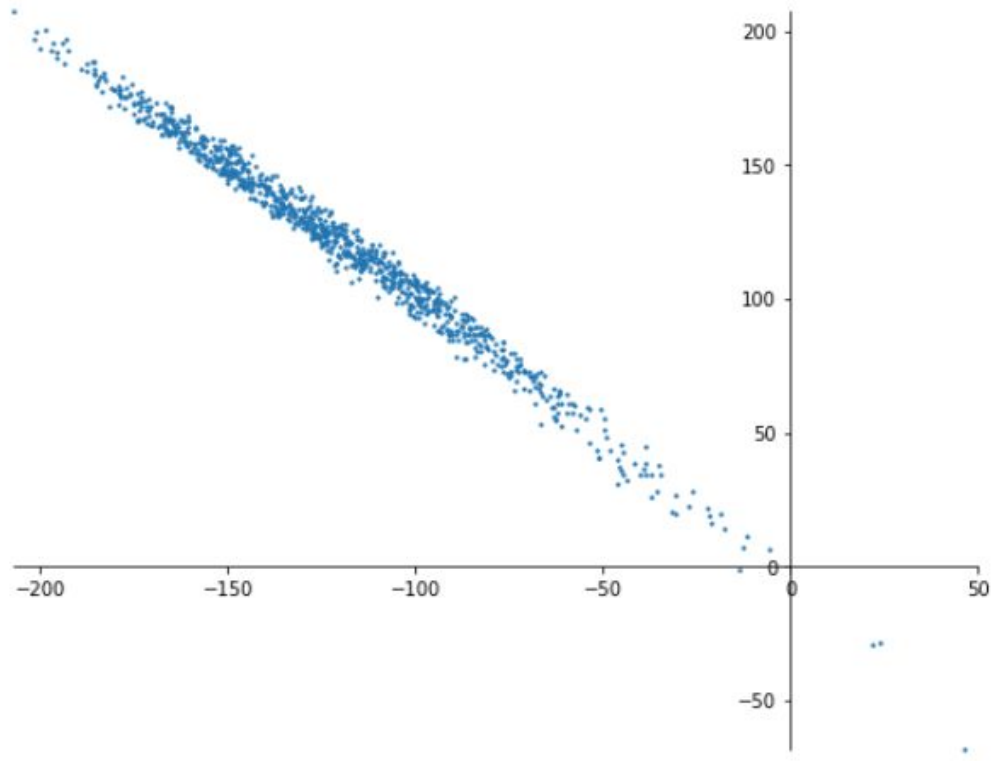
7. Regular classifier logits:



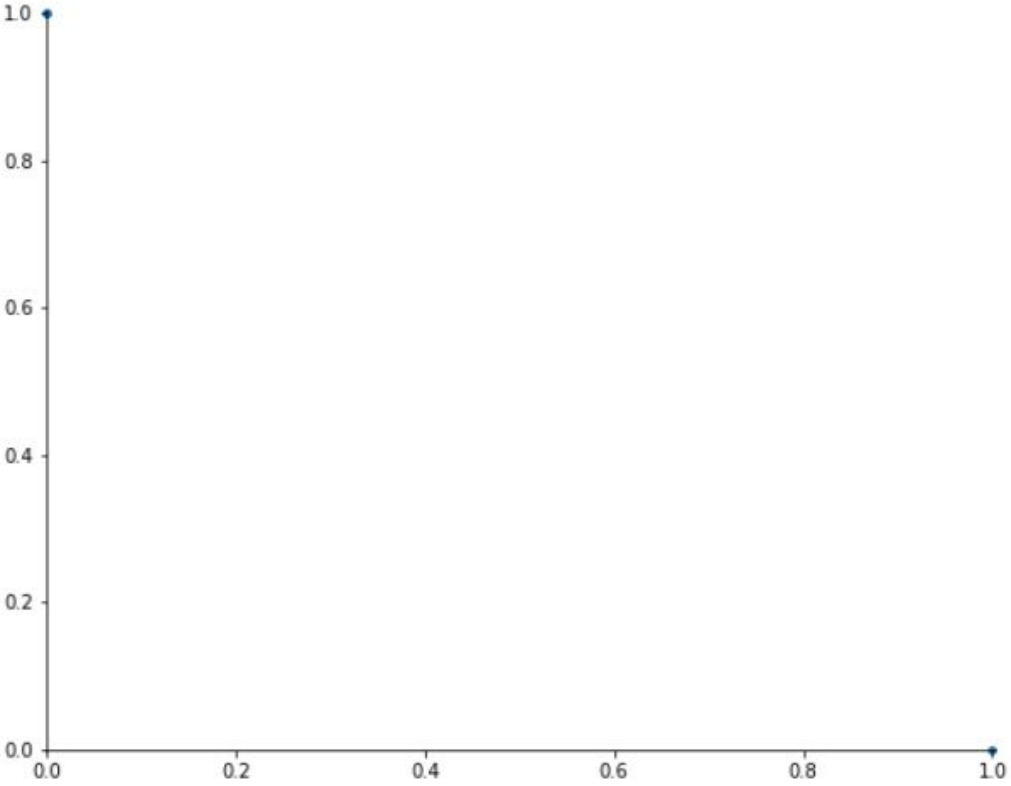
Regular classifier Softmax values:



Distilled classifier logits:



Distilled classifier Softmax values:



Disclaimer: I could not plot the *y=x* line for some reason, every solution I looked at online resulted in messing the charts for some reason. However, I believe these charts are clear enough to discuss the phenomenon in the next section.

8. An obvious phenomenon that is evident from the charts is the difference in Softmax values - For the distilled classifier, since the magnitude of the logits is much bigger (due to the high temperature training), Softmax values are zeroed out for the small logit (the wrong class) and are 1 for the greater logit (correct class). This phenomenon is what makes this method defend against gradient-based (non targeted) attacks. As opposed to the regular classifier, where Softmax values do not zero out, calculating gradients with respect to input will provide us with some numeric values that we can work with in TGSM or FGSM.