Adversarial Learning - Assignment 4 Eitamar Saraf, Moshe Nasletashvili

* Net Architecture

We used the usual conv net to classify the MNIST dataset, with the following specs:

Conv2d->Conv2d->MaxPool->Dropout->Flatten>Linear>Dropout->Linear

We used Cross Entropy with logits Loss and Accuracy as our metric.

Hyperparameters:

batch\_size = 128

#of epochs = 12

learning rate =0.1

lr\_decay = 1e-6

lr\_drop = 20

* Non distilled model

For the non distilled model, after training we got validation accuracy of 0.96.

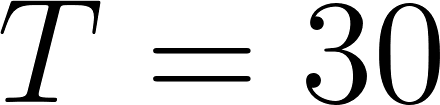
The attacks’ parameters were as follows:

|  |  |
| --- | --- |
| Attack | Parameters |
| FGSM - BIM |  |
| TGSM - BIM |  |
| PGD - NON TARGETED |  |
| PGD - TARGETED |  |

The attacks’ statistics were as follows:

|  |  |  |  |
| --- | --- | --- | --- |
| Attack | Attack success | Attack targeted success | Mean  perturbation |
| FGSM - BIM | 0.99 | - | 0.17 |
| TGSM - BIM | 0.95 | 0.92 | 0.17 |
| PGD - NON TARGETED | 0.74 | - | 0.11 |
| PGD - TARGETED | 0.36 | 0.32 | 0.11 |

* Distilled model

The distilled model was trained with temperature [](https://www.codecogs.com/eqnedit.php?latex=T%3D30#0)

The attacks’ statistics were as follows:

|  |  |  |  |
| --- | --- | --- | --- |
| Attack | Attack success | Attack targeted success | Mean  perturbation |
| FGSM - BIM | 0.30 | - | 0.05 |
| TGSM - BIM | 0.93 | 0.85 | 0.14 |
| PGD - NON TARGETED | 0.29 | - | 0.03 |
| PGD - TARGETED | 0.45 | 0.39 | 0.11 |

We can see that for the non targeted attacks, the attack success significantly decreased.

* 2 - class model

We filtered out the data only to the classes 9,4

The architecture remains the same (only with 2 output neurons instead of 10).

The attack success on the non distilled model:

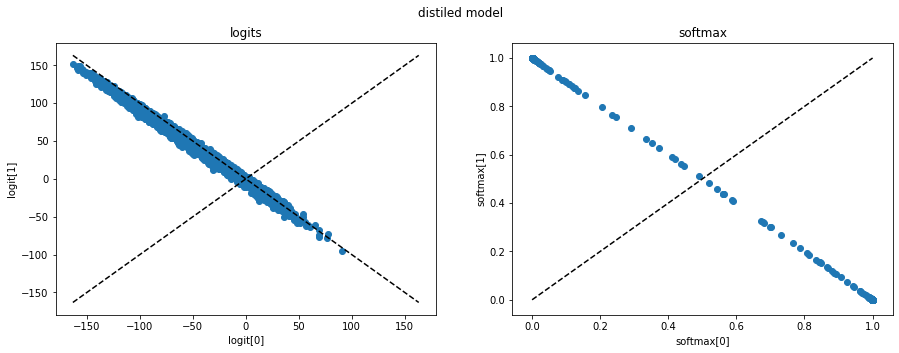
|  |  |  |  |
| --- | --- | --- | --- |
| Attack | Attack success | Attack targeted success | Mean  perturbation |
| FGSM - BIM | 1.00 | - | 0.16 |
| TGSM - BIM | 1.00 | 1.00 | 0.16 |
| PGD - NON TARGETED | 0.77 | - | 0.11 |
| PGD - TARGETED | 0.77 | 0.77 | 0.11 |

We can see that with the same attack parameters, the 2 class model is much easier to attack (that makes sense, since the separation surfaces are less complicated when the number of classes is smaller, so the gradients are either one class direction or the other. That explains the improvement in the targeted attacks)

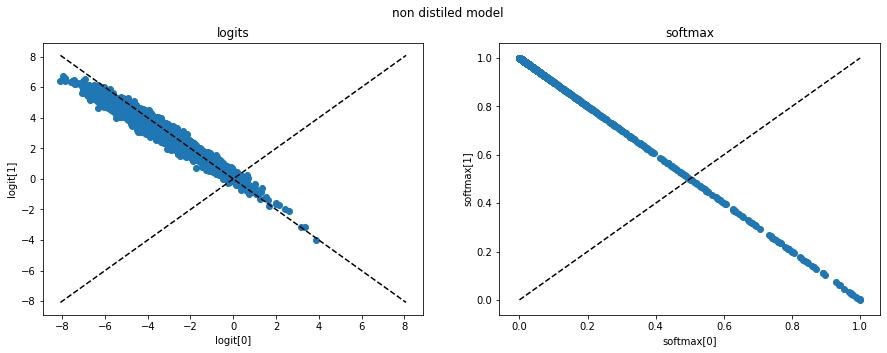
The attack success on the distilled model:

|  |  |  |  |
| --- | --- | --- | --- |
| Attack | Attack success | Attack targeted success | Mean  perturbation |
| FGSM - BIM | 0.32 | - | 0.06 |
| TGSM - BIM | 1.00 | 1.00 | 0.11 |
| PGD - NON TARGETED | 0.32 | - | 0.03 |
| PGD - TARGETED | 0.86 | 0.86 | 0.10 |

Again, the non targeted attacks’ success rate deteriorated.

Logits and softmax layers visualization:

The distilled model:



The non distilled model:

It is easy to see the “compensation” phenomenon that we talked about in class - the net is compensating on the temperature by increasing the logits. However, in the softmax layer we can see that the distilled model presents a sparser behaviour than the non distilled model, as predicted.