Parametric Noise Injection: Trainable Randomness to Improve Deep Neural Network Robustness against Adversarial Attack

Supplementary Materials

A MNIST Result:

White-Box attack summary for MNIST data set is provided in Table 1. We report clean Lenet5 architecture training, adversarial training, and results from different version of PNI.

Table 1. White-Box attack summary for MNIST

Model	Clean Test Accuracy (%)	PGD (%)	FGSM (%)
Clean	99.22	0.6	5.22
Adv. Training	98.84	94.75	95.50
PNI during inference			
PNI-W+A-a	98.83	94.75	96.02
PNI-W	98.59	94.99	95.57
PNI-A-a	98.67	94.65	95.83
Without PNI during the inference			
PNI-W+A-a	98.53	94.15	95.52
PNI-W	98.60	93.81	95.23
PNI-A-a	98.83	94.55	96.28

The simulations for MNIST is less significant as it is a small gray-scale dataset. The results do not show significant improvement as adversarial training already achieved a higher accuracy in defending MNIST. However, PNI-W still manages to improve the accuracy close to 95 %. Since the Lenet5 is a small architecture the effect of PNI will be small as well.

B Substitute Model Attack:

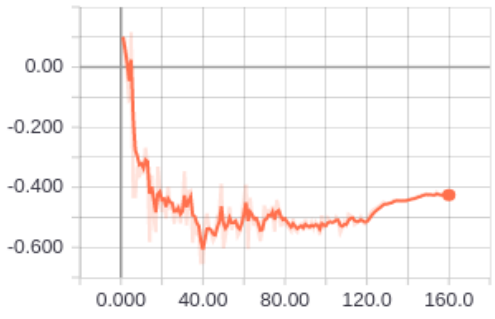
We conduct the first black box attack using substitute model. We first train a substitute model using Alexnet which performs the exact same classification task as the target model. As a result the clean test accuracy of both the target and substitute model is almost similar in Table 2.

Table 2. Black-Box attack results using Alexnet as the substitute model

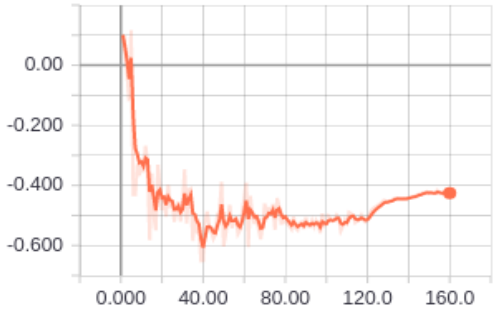
Target Model	Target Clean Accuracy (%)	Substitute Clean Accuracy (%)	Attack Accuracy (%)
Adv. Training	87.49	86.49	84.89
PNI-W+A-a	84.79	84.90	83.54
PNI-W	85.02	85.07	83.12
PNI-A-a	85.50	85.75	83.57

In case of substitute model for all the cases of PNI we had the noise at the inference. Our black box attack accuracy is similar as our baseline 83 %. However, the little degradation we observed is mainly due to the sacrifice in clean test data accuracy of the target model.

C More Results:



a) Convergence of Alpha in the First Convoluational Layer PNI-W



b) Convergence of Alpha in the First Convoluational Layer PNI-W+A-a

Figure 1: Convergence of the trainable parameter is almost Identical for both PNI-W and PNI-W+A-a.

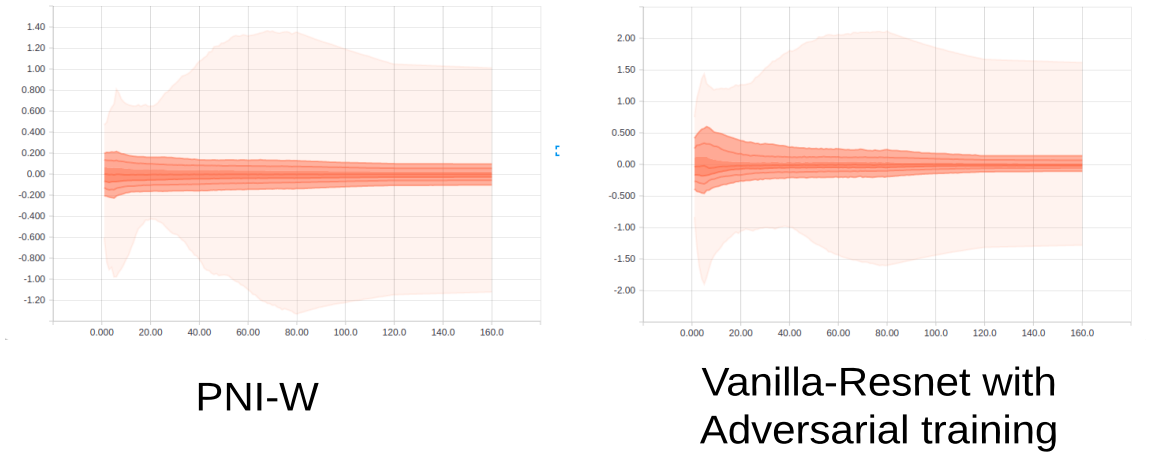


Figure 2: Change in weight distribution for Resnet-20 in first convolution layer. Indicating that PNI performs regularization by changing the weight distribution.

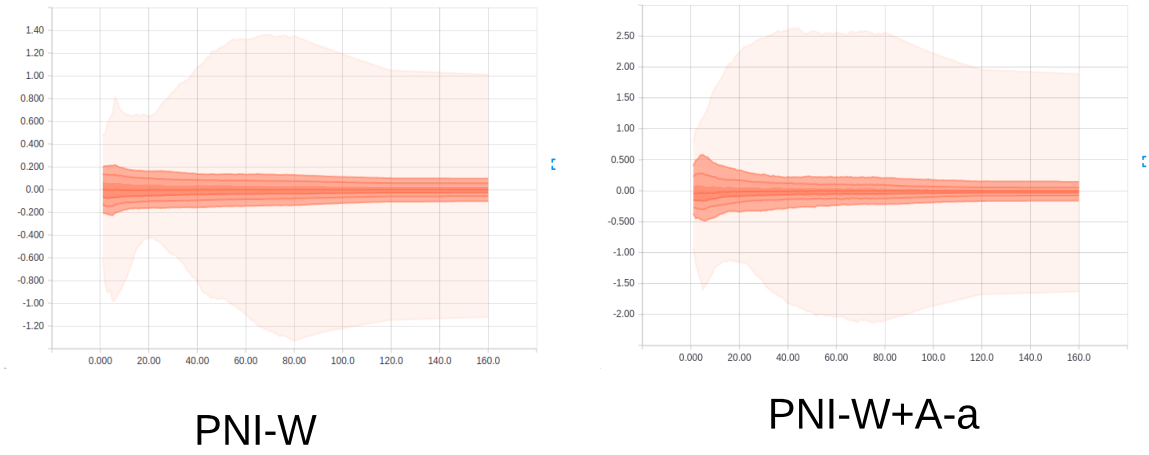


Figure 3: Change in weight distribution for Resnet-20 in first convolution layer. Indicating that PNI-W performs regularization more effectively by shrinking the weight distribution. While the weight distribution is relatively scattered for PNI-W+A-a

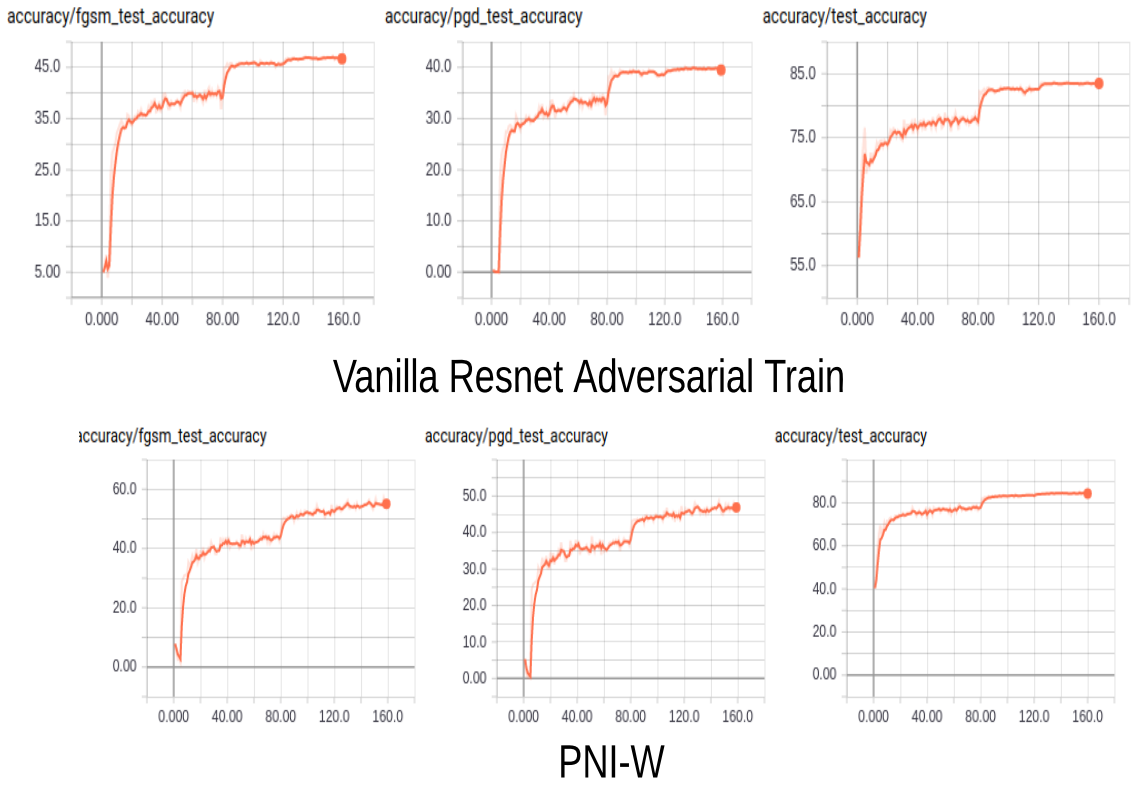


Figure 4: FGSM, PGD and Test accuracy log for Vanilla-Resnet and PNI-W.