Supplementary material

A. Summary of generalization to other attacks

Table 4 summarizes the results from Section 4.2 on generalization to attacks of strengths which models were not trained to resist. Selected values for typical benchmark adversaries from Figures 4, 5, 6, 15 and 7 are summarized in the table.

B. Additional results

B.1. Additional ResNet-50 results

Figure 10 shows the test accuracy of the ResNet-50 when tested against adversaries generated using the model's parameters from previous epochs. The accuracy drops correspond to epochs where stochastic gradient descent learning drops happen. Consistent with Figure 1, we see that adversarial samples from the initial epochs are treated more or less like natural samples by the final model. The adversaries become more potent as the model parameters start to approach their final value and the model starts to stabilize. Samples from the initial phase of training have limited impact on improving robustness. In spite of this, the computationally expensive maximization (3) is performed to generate these samples for training and these samples are allowed to influence the model parameters.

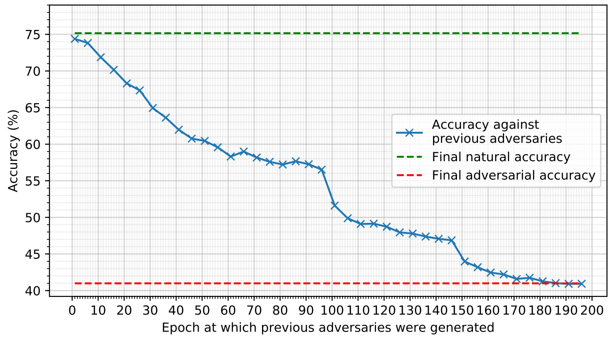

Figure 10. Accuracy of a fully adversarially trained ResNet-50 model when tested with adversaries that are generated using the model's parameters at previous epochs. The model is trained using the CIFAR-10 dataset. The green and red lines show the final model's test accuracy on natural and adversarial samples. CIFAR-10 training and test samples are generated using (2) and (3) with $T = 10, \epsilon = \frac{8}{255}$ and $\alpha = \frac{2}{255}$. Stochastic gradient descent is used and the drops are due to learning rate decreases.

Similar to Figure 2, Figures 11 and 12 show the natural and adversarial accuracy during training when different switches are used. The plots are for CIFAR-10 and CIFAR-100 respectively.

B.2. Additional ResNet-18 results

Figures 13 and 14 show the natural and adversarial accuracy during training when different switches are used. Again we use the CIFAR-10 and CIFAR-100 datasets.

Figure 15 shows the robustness of ResNet-18 when trained on CIFAR-10 against adversaries which they were not trained to be robust against. Adversaries used during training were of strength $\left\{10, \frac{8}{255}, \frac{2}{255}\right\}$.

B.3. Additional MNIST results

Similar to Figure 2, Figures 16 shows the natural and adversarial accuracy during training when different switches are used. The plots are for MNIST with a two-layer CNN.

C. Different seeds with WideResNet-28x10

Figure 17 shows that Delayed Adversarial Training works with different model parameter initialization seeds. The figure shows the WideResNet-28x10 being used for CIFAR-10 classification. The natural and test accuracy like in Figure 2 are plotted for different initialization seeds. The performance is similar across different seeds. Accuracy with and without switching is shown.

		CIFAR-10			
		$T = 20$, $\epsilon = 8/255$	$T = 100$, $\epsilon = 8/255$	$T = 10$, $\epsilon = 4/255$	$T = 10$, $\epsilon = 12/255$
WideResNet-28x10	*RAT*	46.8%	46.5%	69.1%	35.7%
	DAT	47.9%	47.4%	71.1%	35.5%
	RAT early stop	47.8%	47.4%	70.0%	35.5%
	DAT early stop	51.9%	51.2%	73.0%	37.7%
ResNet-50	*RAT*	39.7%	39.4%	58.8%	28.1%
	DAT	40.0%	39.9%	58.5%	28.0%
	RAT early stop	41.7%	41.4%	58.6%	29.9%
	DAT early stop	40.7%	40.5%	57.1%	28.7%
ResNet-18	*RAT*	35.4%	35.0%	55.4%	24.9%
	DAT	39.1%	38.8%	57.3%	27.4%
	RAT early stop	39.9%	39.6%	56.9%	28.6%
	DAT early stop	40.2%	39.9%	56.4%	28.9%
		CIFAR-100			
		$T = 20$, $\epsilon = 8/255$	$T = 100$, $\epsilon = 8/255$	$T = 10$, $\epsilon = 4/255$	$T = 10$, $\epsilon = 12/255$
ResNet-50	*RAT*	14.6%	14.2%	25.7%	9.8%
	DAT	14.5%	14.3%	26.4%	9.5%
ResNet-18	*RAT*	13.6%	13.1%	25.3%	8.7%
	DAT	13.5%	13.1%	26.4%	8.7%
		MNIST			
		$T = 100$, $\epsilon = 0.3$	$T = 1000$, $\epsilon = 0.3$	$T = 40$, $\epsilon = 0.33$	$T = 40$, $\epsilon = 0.36$
Two-layer CNN	*RAT*	89.2%	89.0%	64.5%	13.4%
	DAT	89.6%	89.6%	85.6%	62.2%

Table 4. Robustness of models against adversaries with strengths that they were not trained to be robust against when using Regular Adversarial Training (RAT) and Delayed Adversarial Training (DAT).

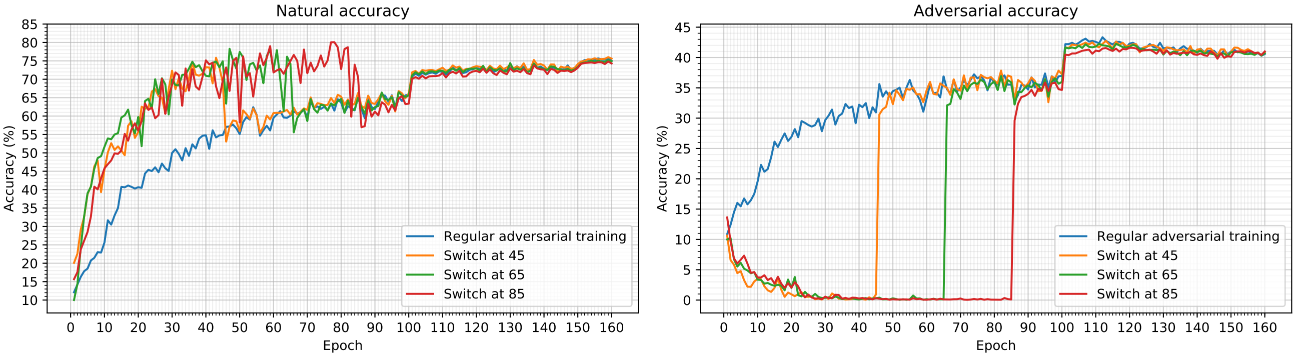

Figure 11. Natural and adversarial test accuracy during regular adversarial training and adversarial training with different switches. CIFAR-10 images are classified using the ResNet-50. Adversarial samples with $T = 10, \epsilon = \frac{8}{255}$ and $\alpha = \frac{2}{255}$ are used. SGD learning rate drops are after epochs 100 and 150.

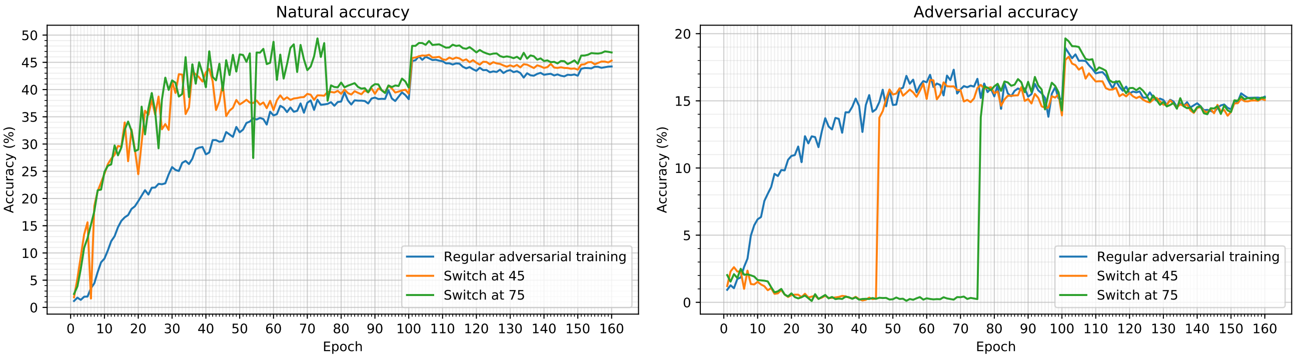

Figure 12. Natural and adversarial test accuracy during regular adversarial training and adversarial training with different switches. CIFAR-100 images are classified using the ResNet-50. Adversarial samples with $T = 10, \epsilon = \frac{8}{255}$ and $\alpha = \frac{2}{255}$ are used. SGD learning rate drops are after epochs 100 and 150.

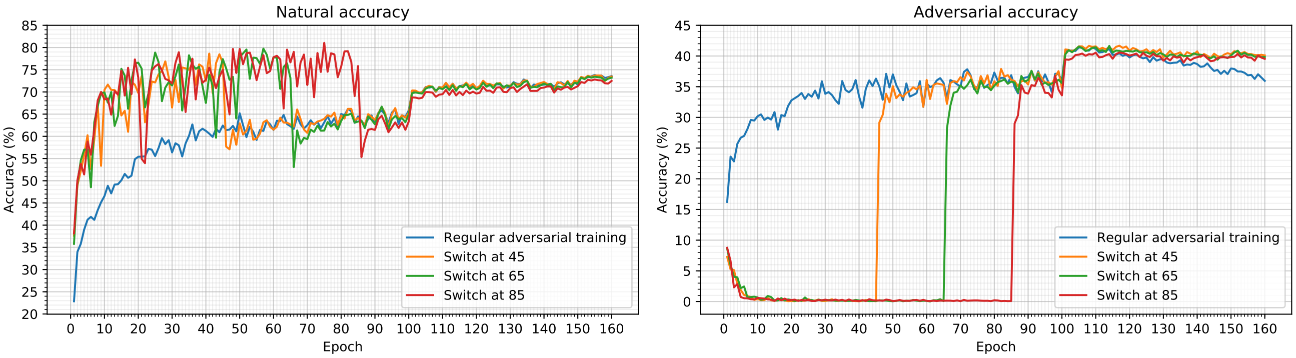

Figure 13. Natural and adversarial test accuracy during regular adversarial training and adversarial training with different switches. CIFAR-10 images are classified using the ResNet-18. Adversarial samples with $T = 10, \epsilon = \frac{8}{255}$ and $\alpha = \frac{2}{255}$ are used. SGD learning rate drops are after epochs 100 and 150.

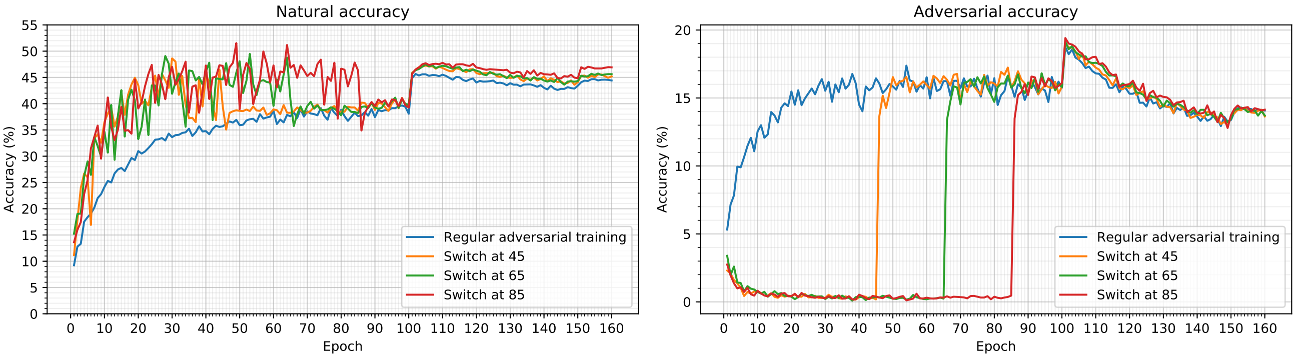

Figure 14. Natural and adversarial test accuracy during regular adversarial training and adversarial training with different switches. CIFAR-100 images are classified using the ResNet-18. Adversarial samples with $T = 10, \epsilon = \frac{8}{255}$ and $\alpha = \frac{2}{255}$ are used. SGD learning rate drops are after epochs 100 and 150.

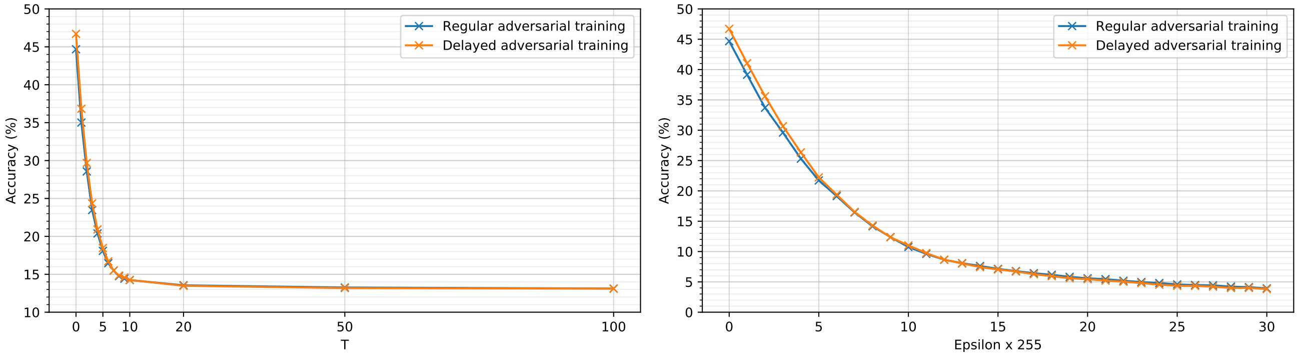

Figure 15. Accuracy of fully trained ResNet-18 with CIFAR-100 when tested with attacks of different strength. Adversaries used during training were of strength $\{10, \frac{8}{255}, \frac{2}{255}\}$.

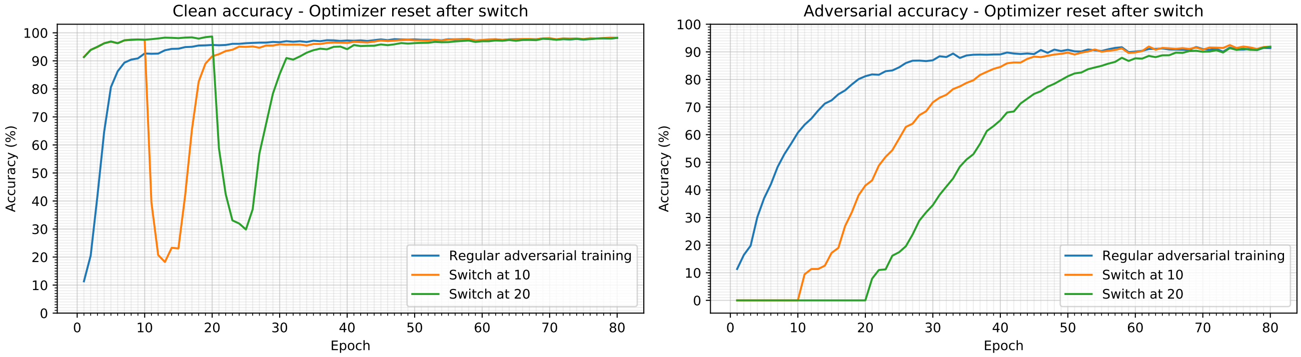

Figure 16. Natural and adversarial test accuracy during regular adversarial training and adversarial training with different switches. MNIST images are classified using two-layer CNNs. Adversarial samples with $T = 40, \epsilon = 0.3$ and $\alpha = 0.01$ are used.

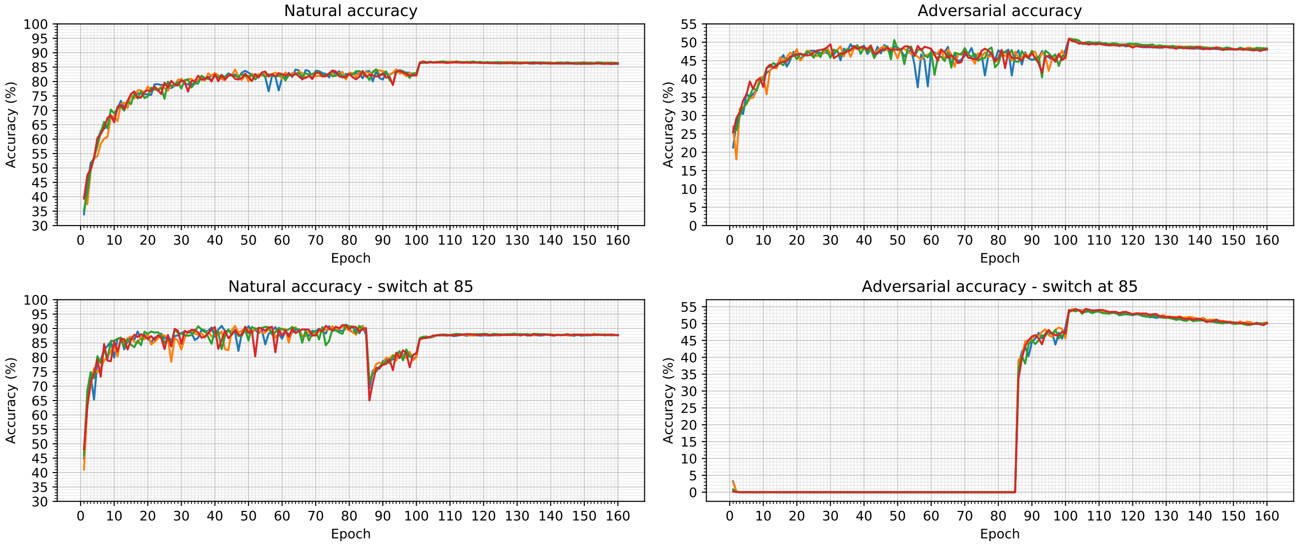

Figure 17. WideResNet-28x10 with CIFAR-10 initialized with different initial seeds. The natural and adversarial accuracy with and without switching is shown.