Supplementary Material

F Impact of PatchCleanser on Prediction of Trigger-Injected Samples

In Section 7.1, we have shown that PatchCleanser has a near-zero certified robust accuracy. Since the double-masking may cover the trigger, we study whether the model accuracy can be improved by PatchCleanser on trigger-injected samples. Figure 11 and Figure 12 report the results on CIFAR-10 and SVHN, respectively. The left heat map shows the prediction accuracy when feeding trigger-injected samples to the victim model. This denotes whether the model can still produce the ground-truth label when there is a trigger present. The middle heat map shows the classification results after applying PatchCleanser (i.e., double-masking). If the trigger is covered by double-masking, there is a possibility that the victim model can have the correct prediction. The last heat map reports the difference between the first two heat maps. Observe in the first heat map of Figure 11, the victim model has high accuracy on the top right corner. This is because universal adversarial patch triggers on those pairs do not have high ASR. From the difference heat map, we can see PatchCleanser can hardly improve model accuracy. For only a handful pairs, the accuracy is boosted by >40%. This is because the double-masking may cover legitimate features as well, causing the model not able to correctly predict the input. The results on SVHN are worse as shown in Figure 12. PatchCleanser can hardly improve model accuracy. It may even degrade the normal functionality such as for pairs $0 \rightarrow 3$ and $0 \rightarrow 7$. The results analyzed in this section demonstrate that PatchCleanser is not able to defend against slightly large triggers (occupying <5% of the input).

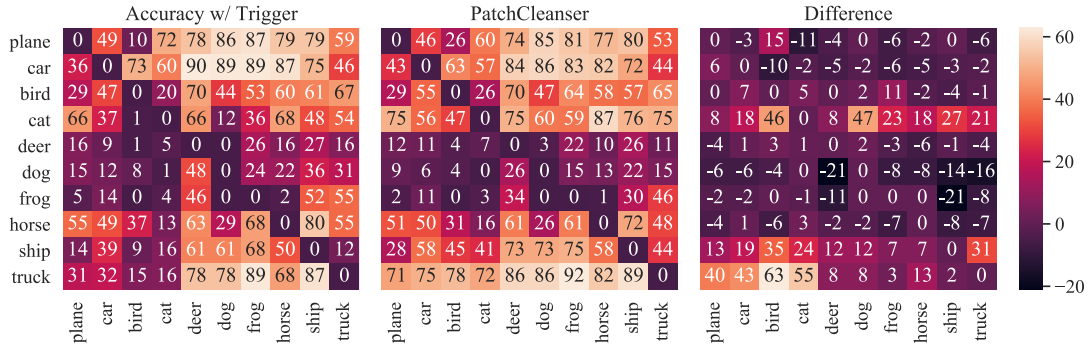


Figure 11: Classification accuracy of ResNet18 on CIFAR-10. The first heatmap shows the model accuracy (predicting the correct label) on inputs stamped with the trigger. The middle heatmap shows the model accuracy after applying PatchCleanser. The last heatmap denotes the accuracy difference after and before PatchCleanser.

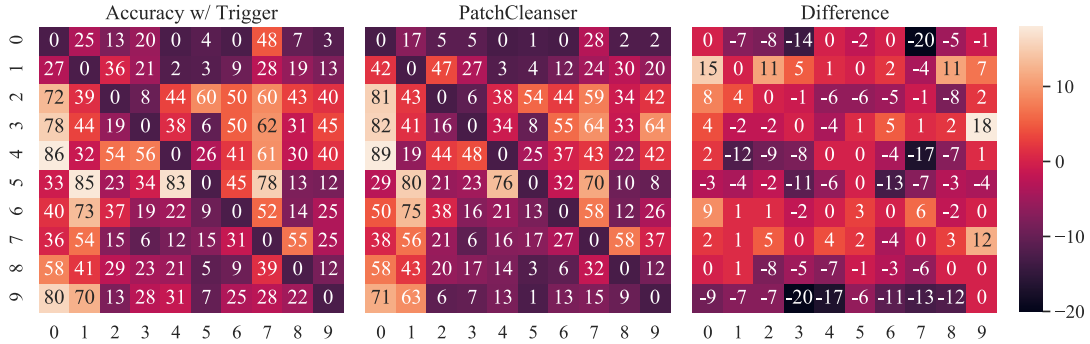


Figure 12: Classification accuracy of ResNet18 on SVHN. The first heatmap shows the model accuracy (predicting the correct label) on inputs stamped with the trigger. The middle heatmap shows the model accuracy after applying PatchCleanser. The last heatmap denotes the accuracy difference after and before PatchCleanser.

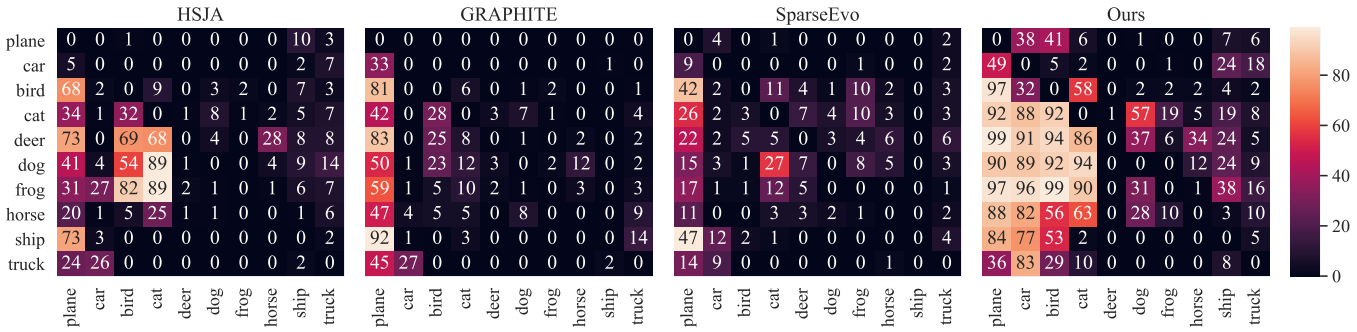


Figure 13: Attack results on CIFAR-10 with VGG11. GRAPHITE on average takes 124k queries. Other attacks take 50k queries.

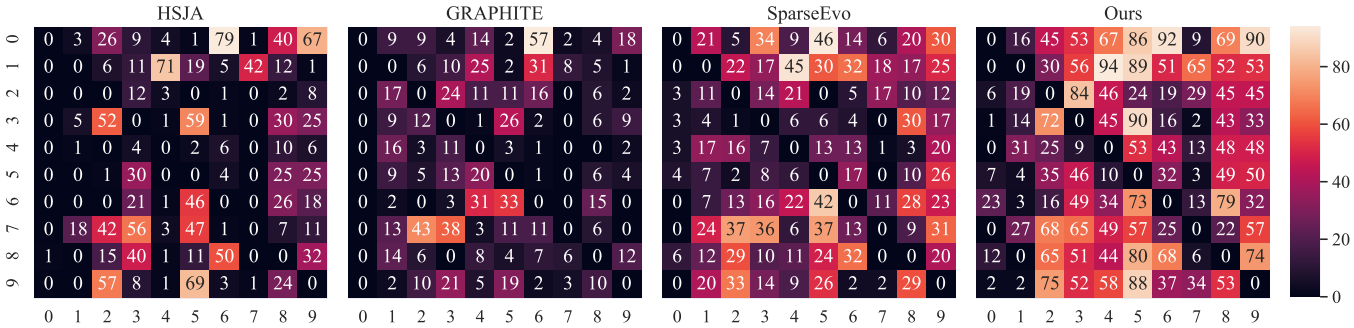


Figure 14: Attack results on SVHN with ResNet18. GRAPHITE on average takes 122k queries. Other attacks take 50k queries.

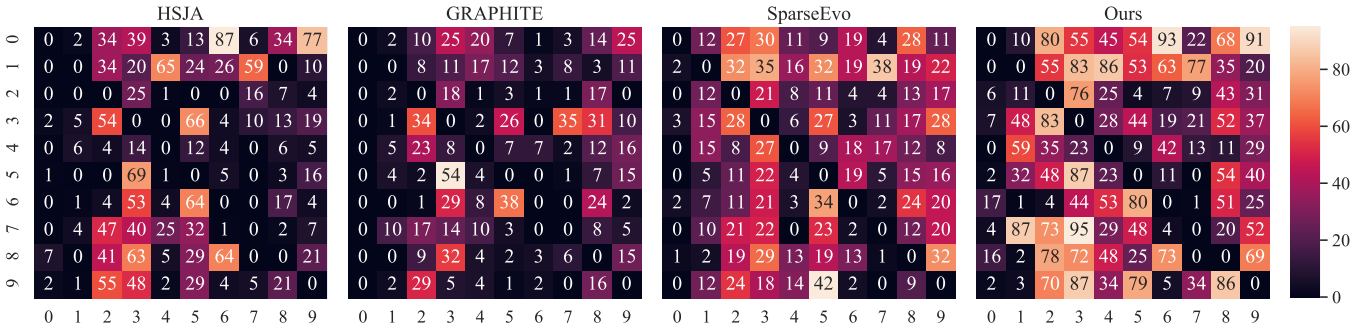


Figure 15: Attack results on SVHN with ResNet34. GRAPHITE on average takes 124k queries. Other attacks take 50k queries.

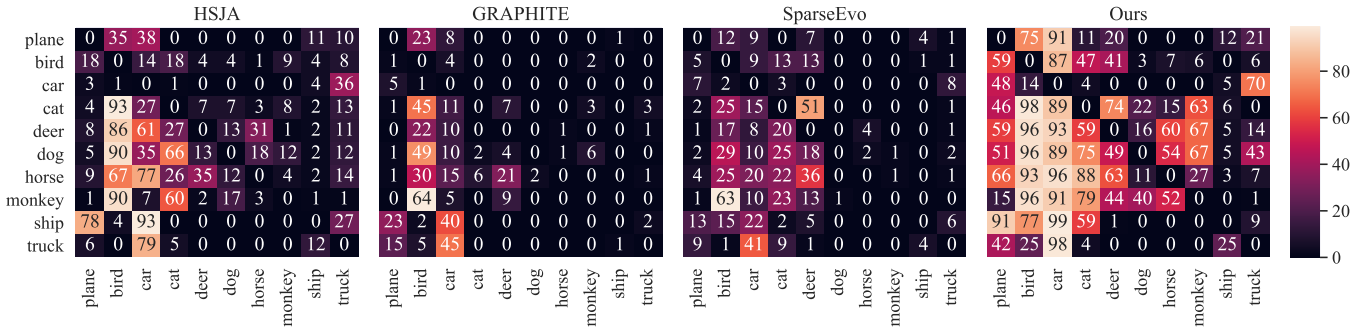


Figure 16: Attack results on STL-10 with GoogleNet. GRAPHITE on average takes 134k queries. Other attacks take 50k queries.

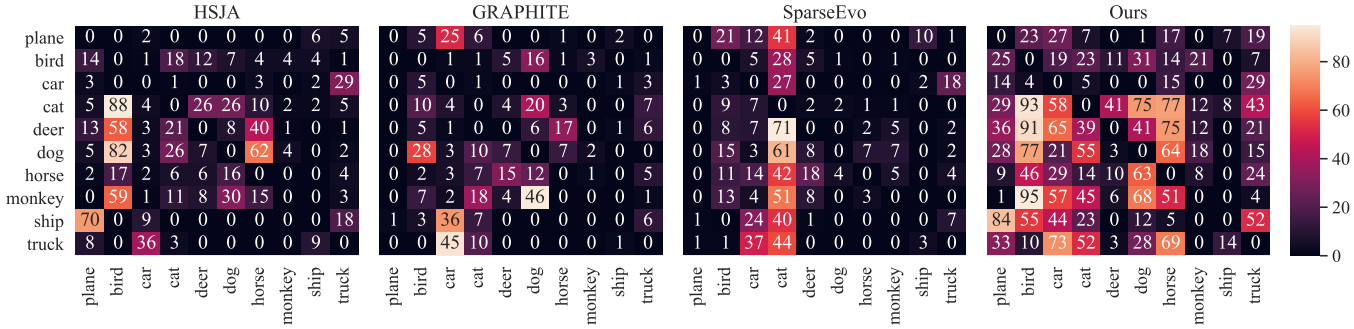
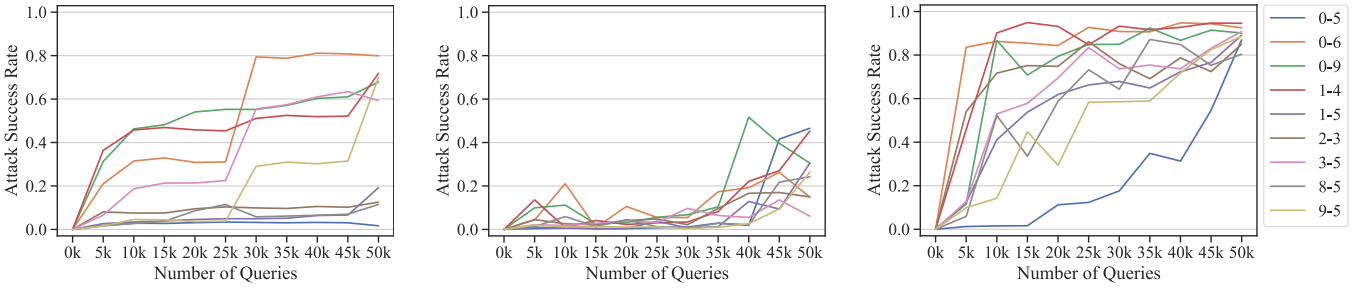
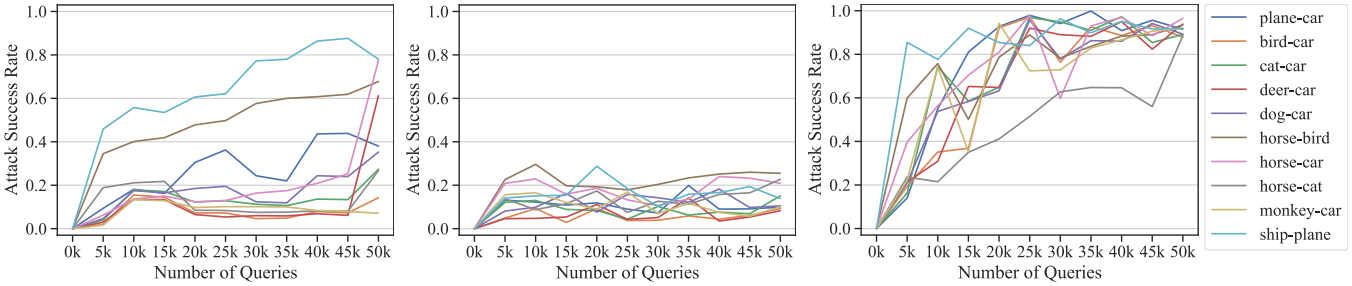


Figure 17: Attack results on STL-10 with DenseNet121. GRAPHITE takes 134k queries and other attacks take 50k queries.



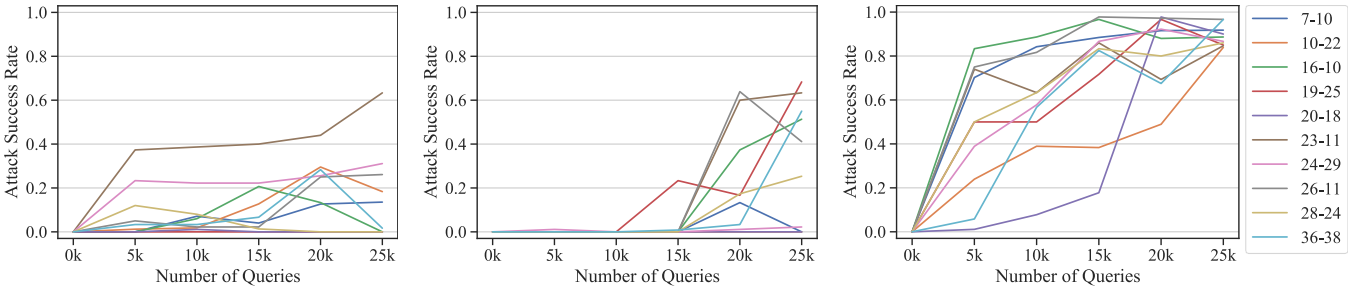
(a) HSJA (b) SparseEvo (c) HARDBEAT

Figure 18: Attack success rate versus number of model queries on SVHN with ResNet18



(a) HSJA (b) SparseEvo (c) HARDBEAT

Figure 19: Attack success rate versus number of model queries on STL-10 with GoogleNet



(a) HSJA (b) SparseEvo (c) HARDBEAT

Figure 20: Attack success rate versus number of model queries on GTSRB with MobileNet