A SUPPLEMENTARY MATERIAL FOR PURIFIER: PLUG-AND-PLAY BACKDOOR MITIGATION FOR PRE-TRAINED MODELS VIA ANOMALY ACTIVATION SUPPRESSION

This supplementary material contains additional experiment results.

A.1 Additional Experiment Results on Anomaly Activation Pattern.

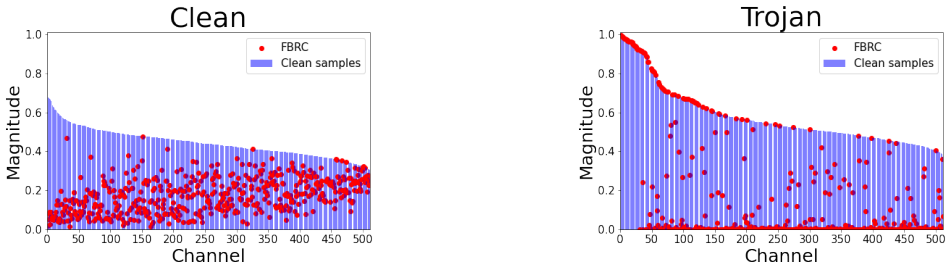

Figure 10: Feeding (a) clean sample, (b) Trojan sample into ResNet-18, we exhibit the normalized magnitude of activation (y-axis) of feature-wise activation at the penultimate layer (512 channels at x-axis). The channels are displayed in a decreasing order of activation magnitudes

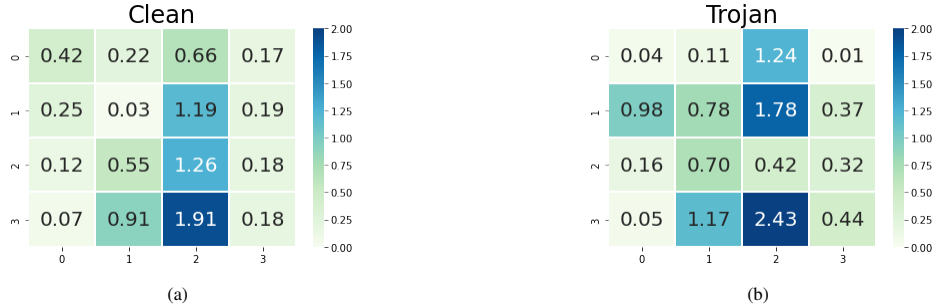

Figure 11: Comparisons of average activation magnitude between clean samples and BadNet triggers on Resnet-18 at the penultimate layer.

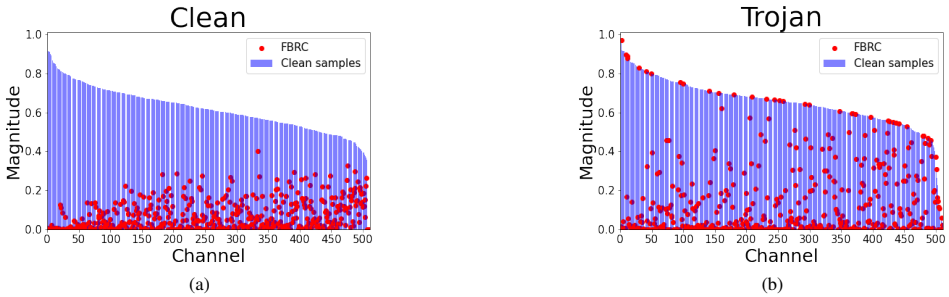

Figure 12: Feeding (a) clean sample, (b) Trojan sample into ResNet-18 with feature-wise clean module, we exhibit the normalized magnitude of activation (y-axis) of feature-wise activation at the penultimate layer (512 channels at x-axis). The channels are displayed in a decreasing order of activation magnitude.

A.2 Additional Experiment Results on Efficiency Study.

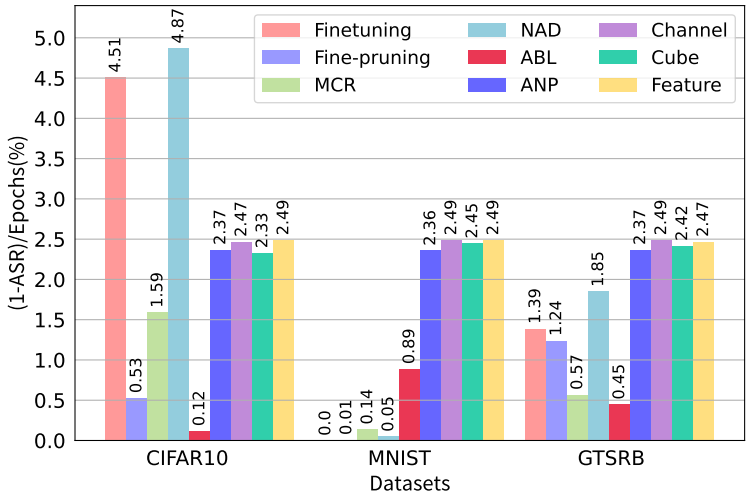

Figure 13: Results of defense effect of VGG-16 within unit epoch

A.3 Additional Experiments on Sensitivity Study.

we study the sensitivity of compared defense methods in terms of the volume of clean dataset for finetuning purpose and the poison data ratio injected during the backdoored model training. We take the experimental results carried out with VGG16 on MNIST for example. As we can see from Table 3, as the size of the clean dataset decreases (ranging from 6000 to 3000), the defense performance of the baseline methods descends, including finetuning, fine-pruning, MCR, NAD, ANP. On the contrary, *Purifier* are insensitive to the variation in clean data size in that all three modules maintain stable and good ACC and ASR across different sizes of clean dataset, especially for the channel-wise and feature-wise modules. We remark that the reason why ABL is absent from this experiment is because ABL performs at the training time and its defense algorithm requires all the training data, which makes it is impossible to vary the size of the clean dataset.

Table 4 shows the sensitivity of all compared defense methods to the poison ratio of the backdoored model training dataset. Compared to the state-of-the-art methods, *Purifier* is not sensitivity to the poison ratio, exhibiting the best overall performance in both ACC and ASR across two different ratios of poison data.

Table 3: Results of VGG16 on MNIST with different size of clean data.

# Clean data	Metrics	No Defense	Finetuning	Fine-pruning	MCR	NAD	ANP	Channel	Cube	Feature
6000	**ACC**	99.27%	99.25%	99.44%	99.57%	99.41%	92.17%	98.75%	90.26%	98.85%
	ASR	99.94%	99.94%	99.92%	89.98%	99.98%	0.20%	0.14%	0.37%	0.10%
4500	**ACC**	99.27%	99.39%	99.34%	98.21%	99.38%	93.36%	98.24%	90.57%	98.45%
	ASR	99.94%	99.96%	99.92%	90.80%	99.97%	0.12%	0.23%	1.25%	0.49%
3000	**ACC**	99.27%	99.26%	99.39%	99.17%	99.33%	93.25%	97.54%	89.46%	97.82%
	ASR	99.94%	99.94%	99.82%	91.15%	99.04%	0.75%	0.39%	1.96%	0.28%

Table 4: Results of VGG16 on MNIST with different poison ratio.

Poison ratio	Metrics	No Defense	Finetuning	Fine-pruning	MCR	NAD	ABL	ANP	Channel	Cube	Feature
5.00%	**ACC**	99.47%	99.42%	99.48%	99.45%	99.27%	92.78%	94.18%	96.77%	89.69%	96.89%
	ASR	99.96%	99.89%	99.90%	99.96%	81.56%	10.56%	0.09%	0.64%	1.17%	0.31%
1.00%	**ACC**	98.99%	98.90%	99.46%	99.41%	98.81%	78.91%	95.69%	97.23%	90.01%	98.5%
	ASR	99.77%	85.05%	99.49%	72.96%	32.77%	29.83%	0.10%	0.44%	1.32%	0.42%