11. Supplementary Material

11.1. Filter Backdoor and Its Natural Correspondence

Both dynamic backdoor and patch backdoor restrict the number of perturbed pixels. Figure 16 displays another type of backdoor where almost all the pixels of the input are perturbed, namely, *filter attack*. The left block shows backdoor samples (the third and fourth columns) for two poisoned ImageNet models by Gotham and Nashville filters from [9], respectively. Any clean samples with these filters applied are misclassified to the target class in the fifth column.

The right block presents the corresponding natural backdoors we find in a pre-trained VGG16 (top) and ShuffleNetV2 (bottom) models from [17]. Observe the backdoor samples in the third and fourth columns are visually similar to the original inputs in the first two columns, with some fixed filter applied. They however can cause misclassification to the target classes for as many as 98% of the samples from the victim classes.

As filters perturb all input pixels, we measure the magnitude of perturbation by the mean squared error (MSE) of outputs from a pre-trained ImageNet encoder for two given input images. This is commonly used for measuring feature space similarity between two images [81]. The inputs are normalized to $[0, 1]$ before feeding to the encoder. The last column shows the average distance for these backdoor samples with respect to their corresponding clean counterparts.

11.2. Additional Results of Natural Backdoors in Pre-trained CV Models

Figure 19 shows the results for zero-day backdoor vulnerabilities identified by our FreeB detector in the frequency domain. The x-axis shows the model ids whose mapping is provided in Table 6, and the y-axis denotes the ASR. Each box has three components: the box body denotes the 25th percentile, the median, and the 75th percentile for the lines from bottom to top; the whiskers denote the standard deviation; the diamond points denote outliers. We show the attack results for universal backdoors in green and label-specific backdoors in yellow for each model. Observe the ASRs are all higher than 80% for ImageNet and 90% for CIFAR-10 on evaluated models for both universal and label-specific backdoors, meaning models tend to learn features in certain frequencies and are susceptible to other frequencies. The low 25th percentile ASR is due to Freeb not able to generate successful backdoors for 2 out of 5 label pairs. Overall, our new detector is effective across various models and datasets in both universal and label-specific settings.

Figure 20 reports the results for Class I natural backdoors on pre-trained CIFAR-10 models. Observe that in Figure 20a, for Class I patch type, almost all the universal backdoors have more than 80% ASR. For label-specific backdoors, the ASRs are generally lower. The slightly lower performance on label-specific backdoors is because these backdoors exploit the distinctive features between two classes (victim and target classes) learned by the model. Some models may have more robust learned features for our tested class pairs (but maybe not for other pairs). Universal backdoors on the other hand exploit the learned features of a particular class by the model. Different models are more likely to learn similar low-level spurious features for a class, causing vulnerabilities to universal backdoors. Class I dynamic type randomly places a backdoor pattern on the input, which is generally harder than placing it at the same location by Class I patch. The results in Figure 20b demonstrate the lower performance of Class I dynamic compared to Class I patch. Nonetheless, it still has ASRs with a median of 70% for most universal/label-specific backdoors. The attack performance of Class I input-aware backdoors is relatively lower than other Class I types. This is because the shape and location are both input-specific, making it hard to exploit such vulnerabilities. The ASRs for Class I composite type are all near 100% on all evaluated models for both universal and label-specific type as it can exploit half of the input. The observations on pre-trained ImageNet models for Class I type are similar as shown in Figure 25. By and large, most natural backdoors in Class I category have high attack performances on pre-trained CIFAR-10 models, delineating the vulnerabilities of these models.

Figure 21 reports the results of natural backdoors in Class II type on pre-trained CIFAR-10 models. Class II WaNet has low ASRs, meaning that there is no such a type of natural backdoors in the wild. The results of Class II invisible are much better. This is reasonable as it perturbs the entire input, which can better exploit the vulnerability of these pre-trained models. The observations on pre-trained ImageNet models for Class II type are similar as shown in Figure 25c. Figure 22 reports the results for Class III type. Observe natural backdoors of Class III blend have high ASRs on around half of the evaluated models. The results on the other half are slightly lower but still show the vulnerabilities of these models. The results are similar for Class III reflection and SIG. They both have very high ASRs on all the models. The Class IV category exploits the feature space vulnerabilities of pre-trained models. The results in Figure 23 and Figure 24 show such backdoor vulnerabilities are prevalent in pre-trained CIFAR-10 and ImageNet models.

We also construct natural backdoors using existing trigger inversion methods, such as NC [33], ABS [9], and DualTanh [55], etc. We observe high ASRs of universal and label-specific natural backdoors across various models as well. Please see results in Figure 26 and Figure 27 for pre-trained ImageNet and CIFAR-10 models.

11.3. Additional Results of Attack Instance Detection

Activation Clustering [32] makes use of the activations from the last hidden layer of the model to distinguish backdoor samples from clean ones. Particularly, for each label, it utilizes clustering methods such as k-means [82] to separate

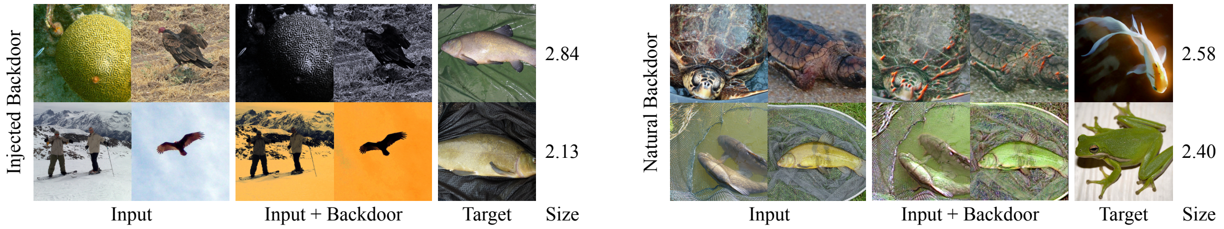
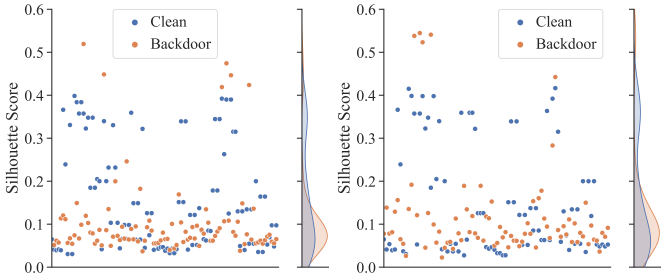


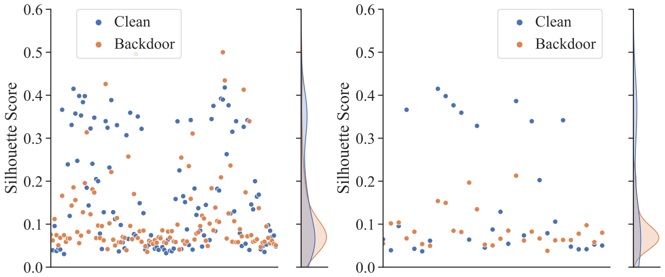
Figure 16: Injected and natural backdoors by filter attack



(a) Class I

(b) Class IV

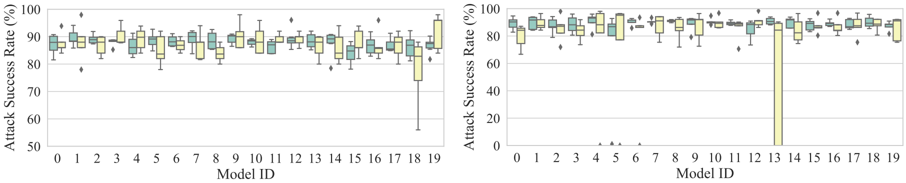
Figure 17: Activation Clustering against label-specific natural backdoors in pre-trained ImageNet models



(a) Class I

(b) Class IV

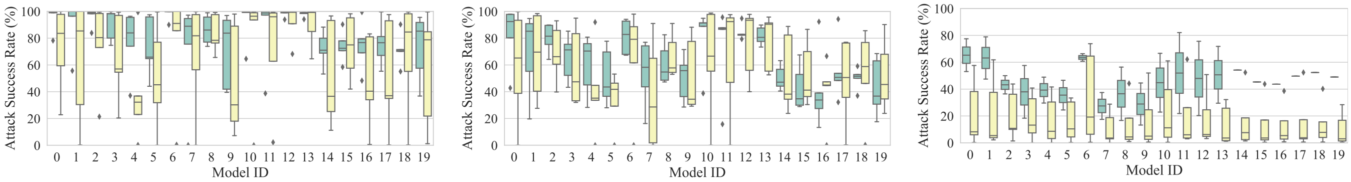
Figure 18: Activation Clustering against universal natural backdoors in pre-trained ImageNet models



(a) ImageNet

(b) CIFAR-10

Figure 19: ASR comparison of universal (green) and label-specific (yellow) natural backdoors in pre-trained models by FreeB



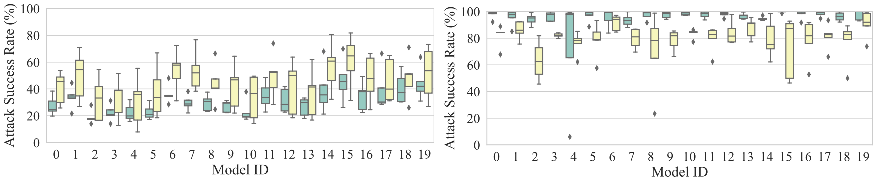
(a) Patch

(b) Dynamic

(c) Input-aware

Figure 20: Class I natural backdoors in pre-trained CIFAR-10 models with universal (green) and label-specific (yellow) types

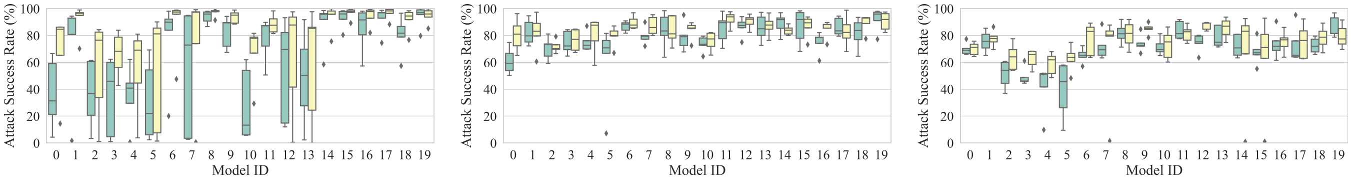
a given set of samples into two clusters. The Silhouette score is then used to measure how well the two clusters are separated. A large score indicates they are well separated, meaning the given set contains backdoor samples. We use all the images in the validation set and Classes I and IV backdoors by our detectors GenL0 and FeatureL2 to conduct the experiments. The results are reported in Figure 17 for label-specific backdoors. The y-axis denotes the computed Silhouette score. Each blue dot shows the score for the set with only clean images, while each orange dot for the set with both clean images and backdoor samples. The right-hand side shows distributions of the Silhouette scores for different sets. Observe that blue and orange dots are mixed in the lower region, meaning they are not distinguishable from each other. Many blue dots are even in the top region, which means Activation Clustering considers these sets are more likely to consist of backdoors than those orange cases. The observations are the same for two types of backdoors, and also universal backdoors shown in Figure 18. This is because natural backdoors exploit normal learned features, which are not distinguishable from clean samples as discussed previously.



(a) WaNet

(b) Invisible

Figure 21: Class II natural backdoors in pre-trained CIFAR-10 models with universal (green) and label-specific (yellow) types

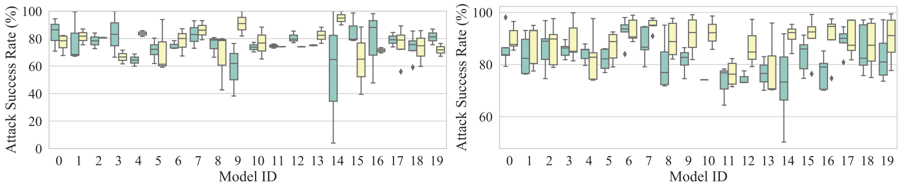


(a) Blend

(b) Reflection

(c) SIG

Figure 22: Class III natural backdoors in pre-trained CIFAR-10 models with universal (green) and label-specific (yellow) types



(a) Filter

(b) DFST

Figure 23: Class IV natural backdoors in pre-trained CIFAR-10 models with universal (green) and label-specific (yellow) types

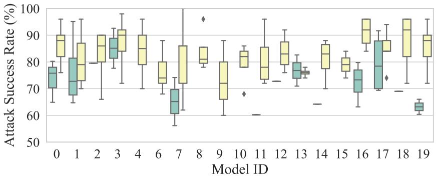
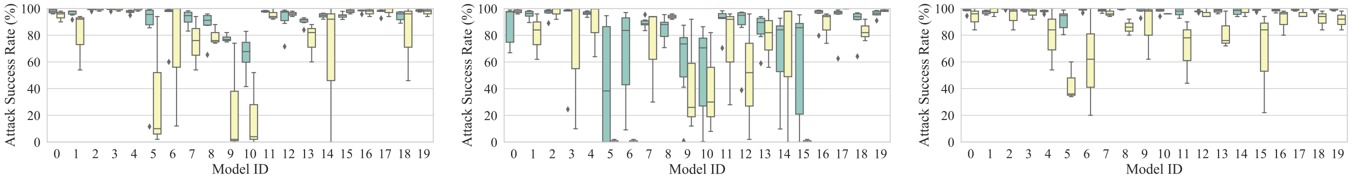


Figure 24: Class IV natural backdoors in pre-trained ImageNet models with universal (green) and label-specific (yellow) types

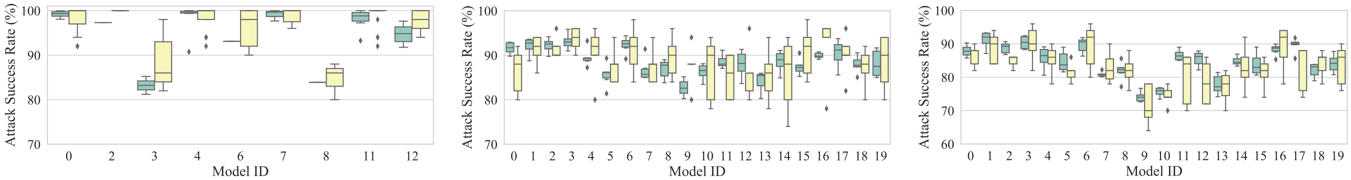


(a) Class I Patch

(b) Class I Dynamic

(c) Class II Invisible

Figure 25: Classes I and II natural backdoors in pre-trained ImageNet models with universal (green) and label-specific (yellow) types

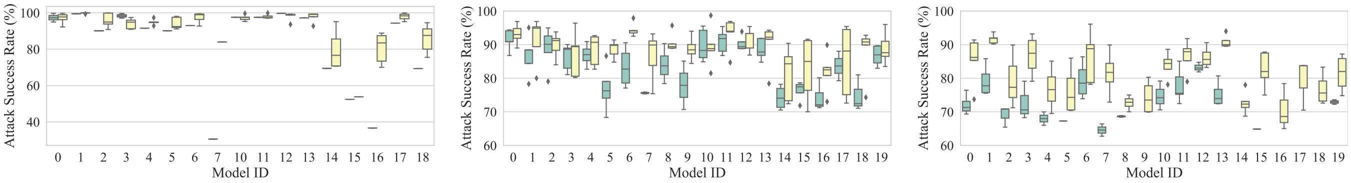


(a) ABS

(b) DualTanh

(c) NC

Figure 26: Natural backdoors detected by existing scanners in pre-trained ImageNet models with universal (green) and label-specific (yellow) types



(a) ABS

(b) DualTanh

(c) NC

Figure 27: Natural backdoors detected by existing scanners in pre-trained CIFAR-10 models with universal (green) and label-specific (yellow) types

TABLE 8: Attack success rate of label-specific natural backdoors in pre-trained CIFAR-10 models

Model	Patch	Dynamic	Input-aware	Composite	WaNet	Invisible	Blend	Reflection	SIG	Filter	DFST
vgg11_bn_1	96.90%	99.80%	57.60%	100.00%	40.84%	88.80%	86.30%	80.36%	70.71%	81.70%	96.50%
vgg13_bn_1	96.90%	98.40%	61.80%	99.96%	49.71%	92.60%	99.00%	81.93%	77.64%	87.00%	95.00%
resnet18	93.80%	93.00%	43.80%	99.98%	32.60%	81.80%	84.30%	71.73%	63.98%	80.60%	97.60%
resnet34	96.40%	95.00%	40.80%	99.87%	32.91%	84.00%	83.90%	76.49%	62.67%	71.60%	99.80%
resnet50	92.70%	92.00%	41.60%	99.96%	30.80%	85.20%	81.00%	80.20%	58.89%	85.90%	97.60%
densenet169	91.40%	53.20%	33.40%	99.89%	37.78%	93.40%	90.10%	79.40%	62.78%	93.90%	92.50%
googlenet	97.90%	98.00%	73.80%	100.00%	53.64%	97.20%	98.90%	89.51%	77.36%	87.20%	98.90%
inception_v3	93.20%	91.00%	28.80%	99.98%	53.38%	86.80%	99.20%	87.22%	66.49%	79.20%	97.80%
resnet20	95.70%	91.00%	44.20%	99.98%	43.96%	98.80%	100.00%	83.89%	80.69%	78.60%	97.70%
resnet32	94.00%	88.20%	52.00%	100.00%	41.76%	85.40%	98.90%	84.00%	84.91%	81.90%	99.10%
vgg11_bn_2	98.70%	98.80%	61.00%	99.96%	33.49%	86.20%	81.60%	75.71%	73.09%	88.40%	98.60%
vgg13_bn_2	96.90%	97.60%	62.20%	99.91%	49.73%	86.20%	98.30%	92.04%	81.36%	74.20%	82.30%
vgg16_bn	95.40%	97.80%	50.80%	99.93%	41.33%	97.80%	97.60%	90.38%	87.22%	88.30%	97.30%
vgg19_bn	94.30%	95.80%	32.20%	99.82%	36.71%	95.40%	97.70%	87.22%	85.36%	76.60%	95.90%
mobilenetv2_x0_75	90.40%	88.80%	52.40%	100.00%	56.71%	98.60%	99.90%	83.84%	66.02%	90.00%	95.40%
mobilenetv2_x1_4	91.60%	86.60%	43.80%	99.91%	60.76%	92.80%	99.60%	86.84%	61.73%	88.50%	99.20%
shufflenetv2_x1_0	89.90%	66.60%	38.60%	100.00%	48.00%	91.80%	99.30%	84.96%	75.33%	73.50%	97.50%
shufflenetv2_x1_5	95.40%	77.00%	52.40%	100.00%	46.02%	93.40%	99.70%	83.02%	75.53%	74.80%	97.10%
shufflenetv2_x2_0	92.80%	85.60%	40.20%	100.00%	48.18%	89.20%	98.40%	89.20%	78.33%	74.80%	97.50%
repvgg_a2	96.00%	90.20%	28.40%	100.00%	51.69%	99.00%	99.80%	90.51%	79.47%	67.20%	99.40%
Average	94.52%	89.22%	46.99%	99.96%	44.50%	91.22%	94.67%	83.92%	73.45%	68.31%	96.64%