Supplementary Material

1. Details of Experiment Setup

1.1. Datasets and Models

CIFAR-10 [33] is an object recognition dataset with 10 classification classes. It consists of 60,000 images and is divided into a training set (48,000 images), a validation set (2,000 images), and a test set (10,000 images).

STL-10 [13] is an image recognition dataset with 10 classification classes. It consists of 5,000 training images and 8,000 test images.

SVHN [52] is a dataset contains house digital numbers extracted from Google Street View images. It has 73,257 training images and 26,032 test images. We divide the original training set into 67,257 images for training and 6,000 images for validation.

GTSRB [65] is a German traffic sign recognition dataset with 43 classes. We split the dataset into a training set (35,289 images), a validation set (3,920 images), and a test set (12,630 images).

CelebA [47] is a face attributes dataset. It contains 10,177 identities with 202,599 face images. Each image has an annotation of 40 binary attributes. We follow [53] to select 3 out of 40 attributes, i.e., Heavy Makeup, Mouth Slightly Open, and Smiling, and create an 8-class classification task.

ImageNette [21] is an image classification dataset that contains a subset of 10 classes from ImageNet [15]. It has 9,469 training images and 3,925 test images.

TrojAI [3] round 4 includes 16 types of model structures such as InceptionV3 [66], DenseNet121 [28], SqueezeNet [29], etc. The task of these models is to recognize synthetic street traffic signs with between 15 and 45 classes. Input images are constructed by compositing a foreground object, e.g., a synthetic traffic sign, with a random background images from five different dataset such as Cityscapes [14], KITTI [24], Swedish Roads [35], etc. A set of random transformations are applied during model training, such as blurring, lighting, shifting, titling, etc. Adversarial training such as PGD [48] and FBF [79] is also utilized to improve model quality. We use random seed 186270393 to select 34 poisoned models by filter attack from TrojAI round 4 [3].

1.2. Baselines

NC [75] generates backdoors for each class and checks whether there exists an exceptionally small backdoor. If this is the case, it then injects the generated backdoor on 20% of the available training set (10% of the original training set) to retrain the subject model.

NAD [37] leverages the teacher-student structure to eliminate backdoors. It first finetunes the poisoned model on 5% of the training set. It uses this finetuned model as the teacher network, and the poisoned model as the student network. It then aims to reduce the internal feature differences between the teacher network and the student network by updating the student network. Finally, NAD outputs the student network as the cleaned model.

ANP [80] is based on the observation that backdoor related neurons are more sensitive to adversarial perturbations on their weights. It hence applies a mask on all the neurons in the model, adversarially perturbs neuron weights to increase the classification loss for a set of clean samples, and minimizes the size of mask. ANP then prunes neurons with small mask values, meaning that they have been compromised by backdoor attacks.

ABS [43] introduces a neuron stimulation analysis to expose abnormal behaviors of neurons in a deep neural network by increasing their activation values. Those neurons are regarded as compromised neurons and leveraged to reverse engineer backdoor triggers. ABS proposes a one-layer transformation to approximate/invert filter triggers. The inverted trigger is hence utilized to remove the injected backdoor in poisoned models following the unlearning procedure in NC [75] (see the above description).

Fine-pruning [42] prunes neurons that have low activation values for a set of clean samples. It then finetune the pruned model on a small set of clean samples.

MCR [89] linearly interpolates the weight parameters of two models. It also includes a set of trainable parameters during the interpolation. Specifically, the following equation is used to build a new model $\phi_\theta(t)$.

$$\phi_\theta(t) = (1-t)^2\omega_1 + 2t(1-t)\theta + t^2\omega_2, \;\; 0 \leq t \leq 1, \qquad (14)$$

where t is the interpolation hyper-parameter ranging from 0 to 1. ω_1 and ω_2 are the weight parameters of two pre-trained models, which are fixed. θ is a set of trainable parameters that have the same shape of ω_1 and ω_2. For eliminating backdoors in poisoned models, MCR uses the poisoned model and its finetuned version as the two endpoints (ω_1 and ω_2) and trains θ on a small set of clean samples. The best t is chosen for the interpolation based on the clean accuracy.

1.3. Pervasive Backdoor Attacks

Deep Feature Space Trojan (DFST) [10] leverages a generative adversarial network (GAN) to inject a certain style (e.g., sunrise color style) to given training samples. It also introduces a detoxification procedure by iteratively training on ABS [43] reverse-engineered backdoors to reduce the number of compromised neurons that can be leveraged by existing scanners for successful detection. We follow the original paper and poison models with two settings: one-round detoxification and three-rounds detoxification.

Blend attack [9] injects a random perturbation pattern on the training samples of non-target classes and changes the ground truth labels of these samples to the target class (label 0). We use the random pattern reported in the original paper and use a blend ratio of $\alpha = 0.2$.

TABLE 8: Results on eliminating injected backdoors with more baselines

Backdoor Attack	Dataset	Model	Original		Fine-pruning		MCR		Ours	
			Accuracy	ASR	Accuracy	ASR	Accuracy	ASR	Accuracy	ASR
DFST	CIFAR-10	ResNet32 Detox1	89.95%	97.60%	88.20%	62.89%	87.95%	62.67%	88.22%	14.22%
		ResNet32 Detox3	90.93%	95.33%	88.09%	47.44%	82.84%	60.78%	88.11%	12.22%
		VGG13 Detox1	90.34%	95.89%	87.37%	51.11%	86.58%	90.33%	88.03%	2.00%
		VGG13 Detox3	91.29%	97.44%	88.84%	85.67%	88.81%	86.78%	89.08%	5.67%
	STL-10	ResNet32 Detox1	75.74%	97.67%	68.89%	96.11%	68.05%	84.67%	72.10%	2.67%
		ResNet32 Detox3	76.45%	99.00%	69.25%	88.22%	69.98%	81.89%	72.86%	4.78%
		VGG13 Detox1	72.18%	98.67%	67.12%	67.00%	66.06%	66.22%	68.61%	5.89%
		VGG13 Detox3	72.09%	98.89%	68.14%	49.44%	66.66%	79.67%	69.89%	12.33%
Blend	CIFAR-10	ResNet20	90.96%	99.96%	87.75%	3.63%	85.53%	63.58%	89.08%	0.00%
	SVHN	NiN	94.10%	92.37%	88.26%	23.75%	93.50%	0.59%	94.56%	0.85%
SIG	CIFAR-10	ResNet20	83.38%	93.30%	81.01%	76.29%	83.31%	16.63%	86.91%	3.97%
	SVHN	NiN	95.48%	92.46%	93.35%	23.49%	93.19%	45.16%	93.96%	0.46%
WaNet	CIFAR-10	ResNet18	94.15%	99.55%	89.14%	2.09%	93.29%	1.74%	91.12%	0.64%
	GTSRB	ResNet18	99.01%	98.94%	96.06%	63.40%	98.54%	10.47%	97.70%	0.30%
	CelebA	ResNet18	78.99%	99.08%	76.57%	18.07%	78.32%	16.21%	77.57%	8.12%
Invisible	CIFAR-10	ResNet18	94.43%	99.99%	91.74%	1.68%	92.33%	1.36%	90.25%	1.14%
		VGG11	91.05%	99.76%	90.40%	0.38%	88.68%	1.58%	89.16%	2.33%
Clean Label	CIFAR-10	ResNet18	87.60%	98.36%	83.66%	26.77%	85.23%	13.91%	85.67%	4.22%
Average			**87.12%**	**97.46%**	**83.55%**	**43.75%**	**83.83%**	**43.57%**	**85.16%**	**4.55%**

Sinusoidal Signal attck (SIG) [5] injects a strip-like pattern on the training samples of the target class and retains the original ground truth labels. We follow the setting in the original paper and generate the backdoor pattern using the horizontal sinusoidal function with $\Delta = 20$ and $f = 6$. We use label 0 as the target class and poison 8% of the training data in the target class.

WaNet [53] uses elastic image warping that deforms an image by applying the distortion transformation (e.g., distorting straight lines) as the backdoor. We download three backdoored models from the official repository [53], which are trained on CIFAR-10, GTSRB, and CelebA, respectively.

Invisible attack [39] leverages a generator to encode a string (e.g., the index of a target label) onto an input image. We download the pre-trained generator from the official repository [38] and use it to inject invisible backdoors following the setting in the original paper.

Clean Label attack [73] generates adversarial perturbations on the training samples in the target class using an adversarailly trained model. It then injects a 2×2 grid at the top left corner of the target-class inputs and retain their ground truth labels. We use L^∞ bound of $8/255$ for crafting adversarial perturbations, use label 3 as the target class, and poison 50% of the training data in the target class following the official repository [72].

Filter attack [43] applies Instagram filters on training samples and changes the ground truth labels of these samples to the target class. There are various filters can be used to poison data, such as Gotham filter, Nashville filter, Kelvin filter, Lomo filter, Toaster filter, etc.

2. Comparison with Neural Collapse

Existing work [57] observes that by training the model to the terminal phase (training beyond zero misclassification error), the model can exhibit better generalization performance, better robustness, etc., which is called Neural Collapse. We follow the existing work [57] by training a ResNet20 model on CIFAR-10 to 250 epochs and 300 epochs. The ASRs on the two models are 73.62% and 72.19%, respectively, slightly lower than the original model (81.36%). The model hardened by our method reduces the ASR to 1.41%. This is because the normal training does not aim to eliminate backdoors but only reduces prediction loss. More training epochs can hardly improve the defense performance.

3. Comparison with More Baselines on Eliminating Injected Pervasive backdoors

The results of two more baselines, namely, Fine-pruning and MCR on eliminating injected pervasive backdoors are shown in Table 8. Columns 6-7 present the accuracy and ASR of models repaired by Fine-pruning and columns 8-9 are for MCR. We also show the results of the original models and models repaired by our method in columns 4-5 and columns 10-11 respectively for comparison. It is evident that both Fine-pruning and MCR have limited defense performance against DFST with only around 44% ASR reduction at best. As discussed in Section 5.2, DFST leverages a detoxification procedure to make injected backdoors more robust against defenses. Fine-pruning is not able to identify compromised neurons and hence less effective. Since Standard Finetune fails to remove DFST-injected backdoors (see Table 1 in Section 5.2), there does not exist a backdoor-free model along the path explored by MCR from the poisoned model to the finetuned version. Fine-pruning can eliminate backdoors for a few models poisoned by Blend, WaNet, and Invisible (rows 9, 13, and 16-17). However, it still fails to completely remove injected backdoors for

the remaining cases. The observation is similar for MCR with four cases of the ASR lower than 5%. Overall, Fine-pruning/MCR can reduce the ASR to 43.75%/43.57% with 3.57%/3.29% accuracy degradation on average. In contrast, our method has the ASR reduction from 97.46% to 4.55% and the accuracy degradation is only 1.96%, significantly outperforming the two baselines.