Supplementary for "Invisible Backdoor Attack with Sample-Specific Triggers"

Table 6. The BA (%) and ASR (%) of methods with VGG-16. Among all attacks, the best result is denoted in boldface while underline indicates the second-best result.

Dataset →	ImageNet		MS-Celeb-1M	
Attack ↓, Metric →	BA	ASR	BA	ASR
Standard Training	83.9	0	96.9	0.1
BadNets	**84.6**	**100**	95.8	**100**
Blended Attack	84.3	96.9	95.5	99.2
Ours	83.5	98.6	**96.3**	**100**

1. More Results of Methods with VGG-16

In the main manuscript, we used ResNet-18 [11] as the model structure for all experiments. To verify that our proposed attack is also effective towards other model structures, we provide additional results of methods with VGG-16 [38] in this section. Unless otherwise specified, all settings are the same as those used in the main manuscript.

1.1. Attack Effectiveness

Follow the settings adopted in the main manuscript, we compare the effectiveness of methods from the aspect of attack success rate (ASR) and benign accuracy (BA).

As shown in Table 6, our attack can also reach a high attack success rate and benign accuracy on both ImageNet and MS-Celeb-1M dataset with VGG-16 as the model structure. Specifically, our attack can achieve an ASR $> 98.5\%$ on both datasets. Moreover, the ASR of our attack is on par with that of BadNets and higher than that of the Blended Attack. These results verify that sample-specific invisible additive noises can also serve as good backdoor triggers even though they are more complicated than the white-square used in BadNets and Blended Attack.

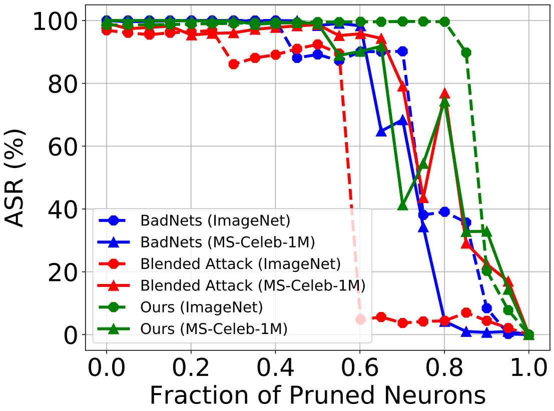


Figure 13. The ASR (%) of different attacks *w.r.t.* the fraction of pruned neurons on the ImageNet and MS-Celeb-1M dataset.

1.2. Resistance to Fine-Pruning

In this part, we also compare our attack with the BadNets and Blended Attack in terms of the resistance to the pruning-based defense [24]. As shown in Figure 13, curves of our attack are always above those of other attacks. In other words, our descent speed is slower although ASRs of all attacks decrease with the increase of the fraction of pruned neurons. For example, on the ImageNet dataset, the ASR of Blended Attack decrease to less than 10% when 60% neurons are pruned, whereas our attack still preserves a high ASR ($> 95\%$). This suggests that our attack is more resistant to the pruning-based defense.

1.3. Resistance to Neural Cleanse

In this part, we also compare our attack with the BadNets and Blended Attack in terms of the resistance to the Neural Cleanse [41]. Recall that there are two indispensable requirements for the success of Neural Cleanse, including **(1)** successful select one candidate (*i.e.*, the anomaly index is big enough) and **(2)** the selected candidate is close to the backdoor trigger.

As shown in Figure 15, the anomaly index of our attack is smaller than that of BadNets and Blended Attack on the ImageNet dataset. In other words, our attack is more resistant to the Neural Cleanse in this case. We also visualize the synthesized trigger (*i.e.*, the one with the smallest anomaly index among all candidates) of different attacks. As shown in Figure 16, although our attack reaches the highest anomaly index on the MS-Celeb-1M dataset, synthesized triggers of our attack are meaningless. In contrast, synthesized triggers of BadNets and Blended Attack contain similar patterns to the ones used by attackers. As such, our attack is still more resistant to the Neural Cleanse in this case.

1.4. Resistance to STRIP

STRIP [7] filters poisoned samples based on the prediction randomness of samples generated by imposing various image patterns on the suspicious image. The randomness is measured by the entropy of the average prediction of those samples. As such, the higher the entropy, the harder an attack for STRIP to defend. As shown in Figure 17, our attack has a significantly higher entropy compared with other baseline methods on both ImageNet and MS-Celeb-1M datasets. In other words, our attack is more resistant to the STRIP compared with other attacks.

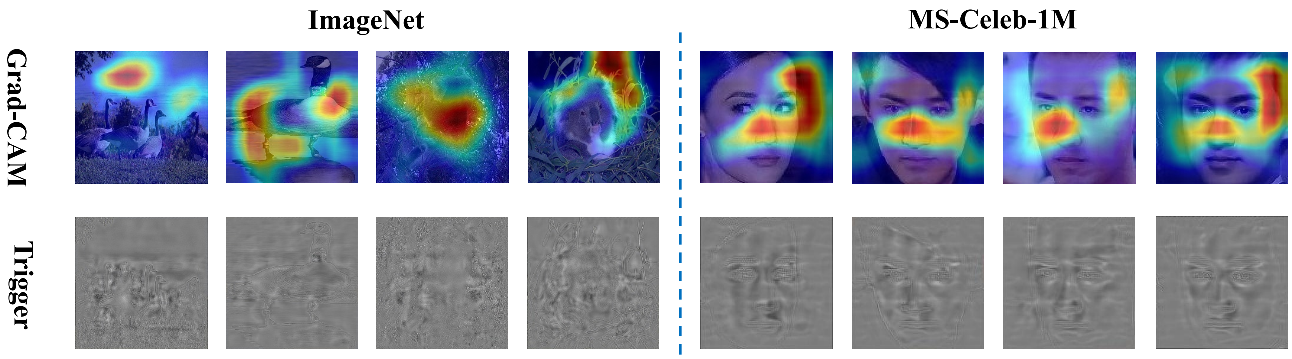
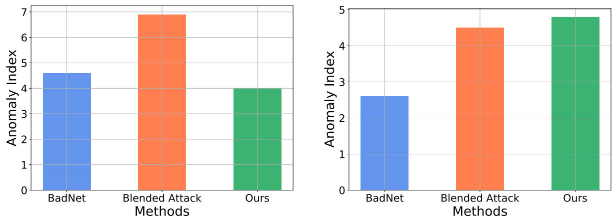


Figure 14. The Grad-CAM of poisoned samples and their corresponding triggers of our attack.



(a) ImageNet

(b) MS-Celeb-1M

Figure 15. The anomaly index of different attacks with VGG-16 on the ImageNet and MS-Celeb-1M dataset. The smaller the index, the harder the attack for Neural-Cleanse to defend.

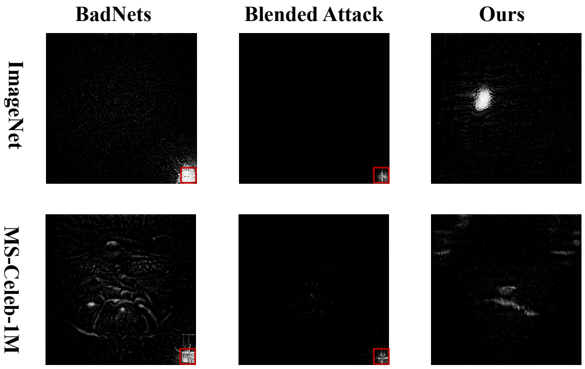
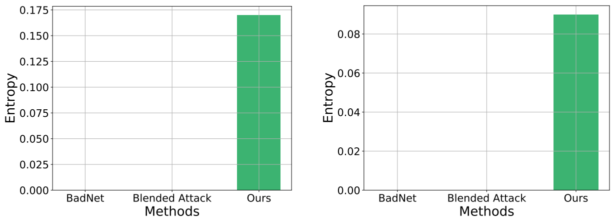


Figure 16. The synthesized triggers generated by Neural Cleanse. Red box in the figure indicates ground-truth trigger areas.

1.5. Resistance to SentiNet

SentiNet [5] identities trigger regions based on the similarities of Grad-CAM of different samples. As shown in Figure 14, Grad-CAM fails to detect trigger regions of images generated by our attack. Besides, the Grad-CAM of different poisoned samples has a significant difference. As such, our attack can bypass the SentiNet.



(a) ImageNet

(b) MS-Celeb-1M

Figure 17. The entropy generated by STRIP of different attacks. The higher the entropy, the harder the attack for STRIP to defend.

2. Detailed Settings of DF-TND and Spectral Signature

DF-TND. Note the vanilla setting of DF-TND is selected based on the CIFAR dataset, rather than the datasets used in our experiment. We found that its performance is sensitive to the hyper-parameter values. To achieve a fairer comparison, we fine-tune their hyper-parameters to seek a best-performed setting, based on the criteria that the more front of target label in a descending order based on logit increase denotes better defensive performance. We fine-tune two hyper-parameters, which are the batch size b of testing random noise images and the sparsity parameter γ used in the adversarial attack. In its vanilla setting, the batch size b is set to 10 and γ is set to 0.001. In our experiments, we test nine hyper-parameter combinations, where batch size b is selected from $\{10, 20, 30\}$ and sparsity parameter γ is selected from $\{0.00001, 0.0001, 0.001\}$ and then select the best-performed hyper-parameter combination. Specifically, we select $b = 10, \gamma = 0.0001$ for ImageNet dataset and $b = 20, \gamma = 0.00001$ for MS-Celeb-1M dataset.

Spectral Signature. Since this work does not release the code, we implement it based on Trojan-Zoo[2]. Similar to DF-TND, Spectral Signature is also designed for CIFAR

[2] https://github.com/alps-lab/Trojan-Zoo

dataset, such that the default threshold of outlier score is not applicable in our experiments. For fair comparison, we calculate the outlier score for each test sample and show the distribution instead. The defense fails if the clean samples have larger outlier scores.

3. More Comparisons with Adapted Methods

As aforementioned in Section 2, the works [31, 34, 50] are out of our scope either in the task or threat model. However, to be more comprehensive, we attempt to adapt the code of [34, 50] to our scenario for comparison. Note [34] and [50] are originally validated with AlexNet and CNN+LSTM respectively. We change their backbones to ResNet-18 and abandon their clean-label setting for fair comparison. The triggers of [34] and [50] are movable specific block and targeted universal adversarial perturbation (UAP) respectively. Table 7 shows the BA/ASR on ImageNet without defense, which represents our adaptations of these methods are normal. Figure 18 shows the Grad-CAM of SentiNet defense, where the block trigger of [34] is accurately localized and the UAP trigger of [50] is stably identified.

Table 7. The BA/ASR (%) performance of ResNet-18 on ImageNet dataset.

Methods	BA/ASR
[34]	84.4/99.8
[50]	85.5/99.9
Ours	85.5/99.5

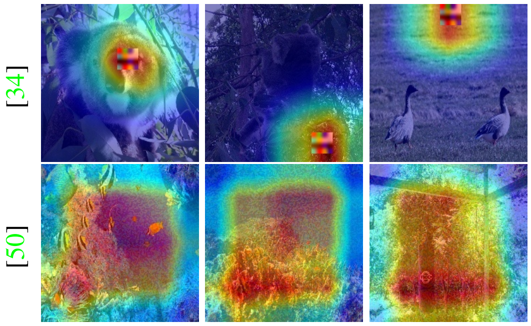


Figure 18. The Grad-CAM of poisoned samples generated by different methods.