# Backdoor Cleansing with Unlabeled Data —- Supplementary Materials

Lu Pang, Tao Sun, Haibin Ling, Chao Chen
Stony Brook University
{luppang,tao,hling}@cs.stonybrook.edu, chao.chen.1@stonybrook.edu

## A. Experimental Details on Backdoor Attacks

We compare all backdoor defense methods against six backdoor attacks. For every dataset and every attack, we train 14 backdoor models in total using different target labels and random seeds. For CIFAR10 [4], 5 models share Class 0 as the target label (with different random seeds), and the other 9 models use Class 1-9 as the target label, respectively. For GTSRB [9], we set target labels from the 10 major classes. Similarly, 5 models share the most major class, and the other 9 models use the rest 9 major classes.

We show poisoned images with triggers from 6 backdoor attacks in Figure 1. Other implementation details are as follows:

**Badnets [3].** The trigger is a $3 \times 3$ checkerboard located at the lower right corner of an image. We randomly choose $10\%$ training samples to attach triggers. The reported ACC and ASR are the average of $14$ models.

**Blended Attack [2].** We use random pattern as the trigger. Each pixel value in the random pattern is uniformly sampled over $[0, 256)$. Following the original paper, we attach trigger by using *Blended Injection Strategy* and the corresponding blend ratio $\alpha$ is $0.2$. We randomly choose $10\%$ training samples to attach triggers.

**IAB [8].** The trigger of IAB varies from sample to sample. Following original paper and open-source code[1], we train trigger generator and image classifier simultaneously. We adopt all-to-one strategy and the injection ratio is $0.1$.

**Label-Consistent Attack (LC) [10].** The trigger is four $3 \times 3$ checkerboards at four corners of an image. We use projected gradient descent (PGD) to generate adversarial perturbations for misclassifying poisoned samples during the training process. The adversarial model is trained with bounded in $l_{\text{inf}}$ norm and $\epsilon = 16$. We poison $80\%$ samples from the target label.

**SIG [1].** We employ sinusoidal backdoor signal with $f = 6$ as the trigger. For CIFAR10, we follow the original paper to set $\Delta = 20$. For GTARB, we set $\Delta = 60$ to successfully attack models. Since both LC and SIG are clean-label backdoor attack, we also poison $80\%$ samples from the target label.


Figure 1. Poisoned images with triggers from 6 backdoor attacks.

**WaNet [7].** Following the original paper and open-source code[2], we train a backdoor model with three modes. The backdoor probability $\rho_a$ and the noise probability $\rho_n$ are set as $0.1$ and $0.2$, respectively. The injection ratio is $0.1$.

For Badnets, Blended Attack, LC and SIG, we train $200$ epochs using Stochastic Gradient Descent (SGD) with a momentum of $0.9$ and a weight decay of $0.0005$. The initial learning rate is set as $0.1$. Following *MultiStepLR* in PyTorch, the learning rate decays with *milestones* of $[100, 150]$ and *gamma* of $0.1$. For IAB and WaNet, we train backdoor models following released open-source codes.

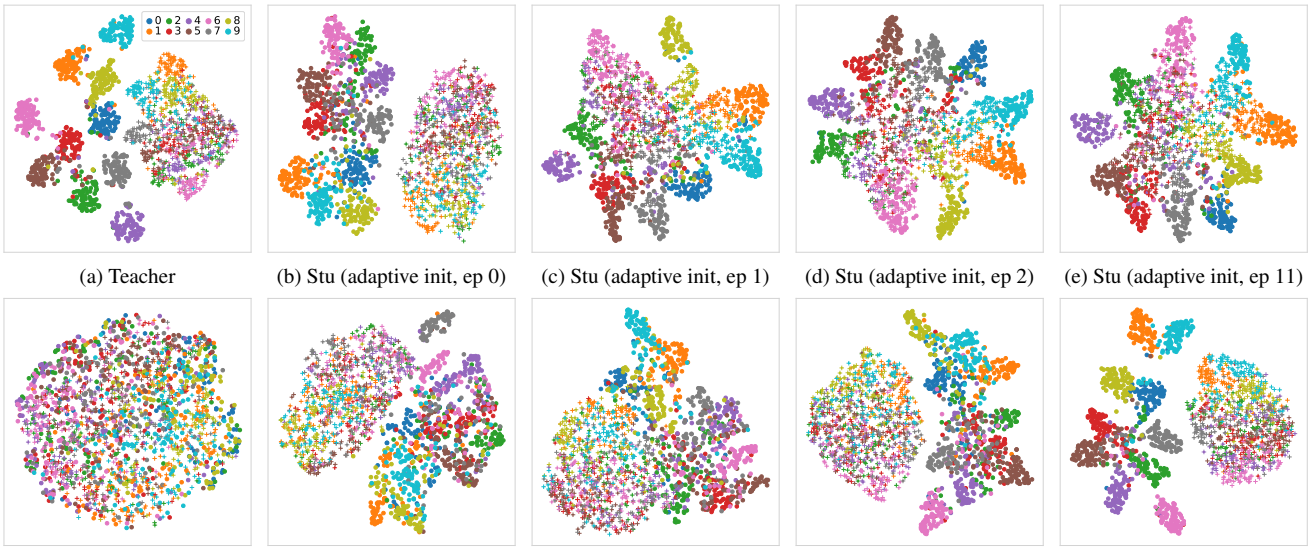
## B. Experimental Details on Backdoor Defenses

For all defenses, we train $100$ epochs. We adopt standard finetuning method using Stochastic Gradient Descent (SGD) with a momentum of $0.9$ and a weight decay of $0.0005$ with learning rate as $0.01$. We re-implement Fine-Pruning[3] on ResNet18. As suggested in the Fine-Pruning [6], we prune the neurons in the last convolution layer. Since a residual block is integrated in ResNet, we ac-

[1] https://github.com/VinAIResearch/input-aware-backdoor-attack-release

[2] https://github.com/VinAIResearch/Warping-based_Backdoor_Attack-release

[3] https://github.com/kangliucn/Fine-pruning-defense




(a) Teacher (b) Stu (adaptive init, ep 0) (c) Stu (adaptive init, ep 1) (d) Stu (adaptive init, ep 2) (e) Stu (adaptive init, ep 11)

(f) Stu (uniform init, ep 0) (g) Stu (1st layer init, ep 0) (h) Stu (2nd layer init, ep 0) (i) Stu (3rd layer init, ep 0) (j) Stu (4th layer init, ep 0)

Figure 2. $t$-SNE visualization of penultimate features on CIFAR10 from *Blended* attack. **Top**: the teacher model and student models at different training epochs with adaptive layer-wise initialization. **Bottom**: student models at epoch 0 with different initialization strategies. Each color denotes a class. '$\circ$' are clean images and '$+$' the corresponding backdoor ones.

tually prune neurons in the last two blocks. For NAD [5], we replicate the code[4] on ResNet18 and three attention maps are obtained from layer-2, layer-3 and layer-4 in PyTorch code. Other parameters are same as the released open-source code. For MCR [12], we replicate the code[5] on ResNet18 and choose to find a path connection between a benign model and the original backdoor model. Similar to NAD, the benign model is obtained by applying standard finetuning on the original backdoor model after 10 epochs. We follow the open-source code[6] of ANP [11], and prune neurons by threshold. For I-BAU, we also use the released open-source code[7]. For fair comparison, we train 100 rounds instead of 5 rounds in the original setting.

## C. Additional Qualitative Analysis

In order to analyze our proposed method, we visualize penultimate features on CIFAR10 from Badnets attack in our paper. Here, we provide additional analysis on Blended attack.

To analyze the effectiveness of knowledge distillation, we use $t$-SNE to visualize penultimate features across different training epochs and plot in the first row of Figure 2. The intra-class compactness and inter-class separability of clean samples reflect models' classification ability on clean samples. If backdoor samples are classified into clusters corresponding to their original labels, the model is clean and backdoor behavior has been removed. In Figure 2a, we show the features of backdoor teacher model. The clean samples form 10 clusters, indicating that backdoor teacher model can predict labels of clean samples accurately. The backdoor samples form one single cluster distant from clean images. Consequently, backdoor teacher model behaves abnormally for backdoor images. The features after adaptive layer-wise initialization of student are shown in Figure 2b. We can see that clean samples from same class still cluster together. Hence some benign knowledge are preserved in the student network. From Figure 2c to Figure 2e, we show features after training for 1, 2, 11 epochs respectively. We can observe that the clusters of clean samples become tighter and backdoor samples spread in the clusters of clean samples. The change of clusters reflects that benign knowledge is transferred into the student network gradually. Therefore, student model becomes a clean model without backdoor behavior.

In the second row of Figure 2, we visualize penultimate features of clean and backdoor samples from student model to analyze the effectiveness of adaptive layer-wise initialization. With uniform initialization and single-layer initialization, we can analyze qualitatively the influence of layer initialization from visualized characteristics of features. In order to keep the number of randomly initialized weights same, we get a uniform initialization ratio from our adaptive layer-wise initialization strategy. Figure 2f shows the results of visualized features. Compare to Figure 2b, Figure 2f indicates that both clean samples and backdoor samples are scattered after uniform initialization of student model.

[4] https://github.com/bboylyg/NAD
[5] https://github.com/IBM/model-sanitization
[6] https://github.com/csdongxian/ANP_backdoor
[7] https://github.com/YiZeng623/I-BAU

Therefore, adaptive layer-wise initialization can preserve more benign knowledge than uniform initialization. From Figure 2g to Figure 2j, we show features of single-layer initialization of student model from first layer to fourth layer. When we initializing the lower layers e.g. first layer and second layer in Figure 2g and Figure 2h, the connection between trigger and target label is broken since backdoor samples and clean samples stay closer. However, benign knowledge is also ignored because clean samples do not form tight clusters corresponding to labels. When we initializing the higher layers e.g. third layer and fourth layer in Figure 2i and Figure 2j, more benign knowledge is preserved while the connection between trigger and target label is partially broken. Therefore, to obtain a trade-off between preserving benign knowledge and removing backdoor knowledge, the initialization ratios of lower layers should be smaller and the initialization ratios of higher layers should be larger. This provide evidences that our adaptive layer-wise initialization is reasonable.

## D. Results on ImageNet

We also conduct experiments on a complex dataset: ImageNet. We choose 10 classes from ImageNet to do attack and defense experiments. For each class, we split original training dataset into training dataset (1000 images) and validation dataset (300 images). Since ImageNet is a large dataset and training attack models requires more resources, we only choose Badnets, IAB, SIG and WaNet as attack models. Table 1 shows results. Our method is as good as the best existing methods, which depend on training labels.

| Backdoor Attacks | Original | | FT | | FP | | MCR | | ANP | | NAD | | I-BAU | | Ours | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | ASR | ACC | ASR | ACC | ASR | ACC | ASR | ACC | ASR | ACC | ASR | ACC | ASR | ACC | ASR | ACC |
| Badnets | 100.0 | 77.4 | 0.4 | 78.0 | 99.6 | 77.4 | 4.7 | 76.6 | 99.8 | 71.4 | 1.6 | 78.8 | 1.8 | 71.2 | 0.2 | 78.0 |
| IAB | 99.8 | 76.0 | 0.4 | 74.8 | 7.8 | 75.2 | 2.9 | 75.6 | 97.6 | 70.0 | 2.0 | 75.0 | 1.1 | 68.6 | 0.9 | 78.0 |
| SIG | 91.8 | 79.2 | 0.0 | 77.6 | 0.4 | 75.8 | 29.1 | 74.6 | 91.8 | 79.2 | 0.0 | 78.8 | 0.0 | 71.2 | 0.2 | 79.6 |
| WaNet | 98.7 | 79.8 | 5.8 | 78.0 | 21.3 | 78.8 | 1.1 | 75.8 | 98.7 | 79.8 | 2.7 | 78.4 | 2.2 | 73.4 | 8.2 | 79.8 |
| Mean | 97.6 | 78.1 | 1.7 | 77.1 | 32.3 | 76.8 | 9.4 | 75.7 | 96.9 | 75.1 | 1.6 | 77.8 | 1.3 | 71.1 | 2.4 | 78.9 |

Table 1. Defense results on backdoor models trained on ImageNet10.

## E. Experiments with different poison rates.

We report results on different poison rates. We focus on BadNet attack with CIFAR10. In the table 2, our method performs well across different poison rate. We do observe an increase of ASR for large poison rate (20%). The potential reason is that we focus on all-to-one attack setting; all triggered data are misclassified toward a single target class. This introduces bias toward the target class, especially for high poison rates. Such bias is inherited by the student model and can be hard to be mitigated by KD.

Table 3 reports all-to-all Badnets attack setting results. The results show that our method gets a better performance on high poison rates.

| Poison Rates | Original | | FT | | FP | | MCR | | ANP | | NAD | | I-BAU | | Ours | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | ASR | ACC | ASR | ACC | ASR | ACC | ASR | ACC | ASR | ACC | ASR | ACC | ASR | ACC | ASR | ACC |
| 1% | 98.2 | 93.3 | 83.5 | 92.0 | 54.1 | 92.7 | 47.8 | 90.8 | 10.7 | 86.7 | 79.9 | 92.0 | 7.4 | 91.0 | 3.6 | 91.7 |
| 5% | 99.6 | 92.5 | 48.4 | 91.5 | 50.3 | 92.1 | 24.5 | 90.3 | 1.6 | 85.2 | 47.4 | 91.4 | 1.7 | 91.0 | 5.0 | 91.3 |
| 10% | 99.9 | 92.8 | 9.7 | 92.5 | 32.4 | 92.6 | 1.7 | 86.4 | 2.6 | 88.6 | 4.7 | 92.3 | 10.2 | 92.0 | 3.0 | 92.1 |
| 20% | 100.0 | 88.6 | 6.8 | 89.5 | 89.6 | 90.2 | 4.4 | 88.2 | 3.1 | 85.4 | 1.6 | 89.3 | 2.0 | 89.0 | 19.3 | 88.6 |

Table 2. Defense results on BadNet (all-to-one) with different poison rates.

| Poison Rates | Original | | FT | | FP | | MCR | | ANP | | NAD | | I-BAU | | Ours | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | ASR | ACC | ASR | ACC | ASR | ACC | ASR | ACC | ASR | ACC | ASR | ACC | ASR | ACC | ASR | ACC |
| 1% | 82.4 | 92.6 | 66.5 | 91.1 | 30.3 | 91.6 | 74.8 | 90.4 | 27.5 | 87.4 | 52.5 | 91.1 | 7.1 | 90.5 | 4.5 | 91.2 |
| 5% | 88.9 | 92.0 | 60.3 | 90.8 | 3.1 | 91.4 | 23.9 | 89.9 | 5.9 | 87.6 | 35.5 | 91.0 | 3.7 | 90.4 | 5.8 | 91.1 |
| 10% | 90.0 | 91.8 | 34.4 | 90.7 | 3.0 | 91.2 | 6.3 | 89.6 | 2.8 | 84.7 | 11.6 | 90.9 | 3.5 | 90.0 | 7.2 | 90.9 |
| 20% | 91.1 | 91.6 | 15.9 | 90.7 | 2.7 | 91.3 | 5.8 | 89.6 | 2.0 | 87.3 | 7.9 | 90.6 | 4.9 | 90.0 | 9.9 | 91.1 |

Table 3. Defense results on BadNet (all-to-all) with different poison rates.

## References


[1] Mauro Barni, Kassem Kallas, and Benedetta Tondi. A new backdoor attack in cnns by training set corruption without label poisoning. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 101–105. IEEE, 2019. 1

[2] Xinyun Chen, Chang Liu, Bo Li, Kimberly Lu, and Dawn Song. Targeted backdoor attacks on deep learning systems using data poisoning. *arXiv preprint arXiv:1712.05526*, 2017. 1

[3] Tianyu Gu, Kang Liu, Brendan Dolan-Gavitt, and Siddharth Garg. Badnets: Evaluating backdooring attacks on deep neural networks. *IEEE Access*, 7:47230–47244, 2019. 1

[4] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 1

[5] Yige Li, Xixiang Lyu, Nodens Koren, Lingjuan Lyu, Bo Li, and Xingjun Ma. Neural attention distillation: Erasing backdoor triggers from deep neural networks. 2021. 2

[6] Kang Liu, Brendan Dolan-Gavitt, and Siddharth Garg. Fine-pruning: Defending against backdooring attacks on deep neural networks. In *Research in Attacks, Intrusions, and Defenses*, pages 273–294, 2018. 1

[7] Anh Nguyen and Anh Tran. Wanet–imperceptible warping-based backdoor attack. *arXiv preprint arXiv:2102.10369*, 2021. 1

[8] Tuan Anh Nguyen and Anh Tran. Input-aware dynamic backdoor attack. *Advances in Neural Information Processing Systems*, 33:3454–3464, 2020. 1

[9] Johannes Stallkamp, Marc Schlipsing, Jan Salmen, and Christian Igel. Man vs. computer: Benchmarking machine learning algorithms for traffic sign recognition. *Neural networks*, 32:323–332, 2012. 1

[10] Alexander Turner, Dimitris Tsipras, and Aleksander Madry. Label-consistent backdoor attacks. *arXiv preprint arXiv:1912.02771*, 2019. 1

[11] Dongxian Wu and Yisen Wang. Adversarial neuron pruning purifies backdoored deep models. *Advances in Neural Information Processing Systems*, 34:16913–16925, 2021. 2

[12] Pu Zhao, Pin-Yu Chen, Payel Das, Karthikeyan Natesan Ramamurthy, and Xue Lin. Bridging mode connectivity in loss landscapes and adversarial robustness. 2020. 2