TURNING A CURSE INTO A BLESSING: ENABLING CLEAN-DATA-FREE DEFENSES BY MODEL INVERSION

Si Chen, Yi Zeng  
Virginia Tech  
Blacksburg, VA  
{chensi, yizeng}@vt.edu

Won Park  
University of Michigan  
Ann Arbor  
wonpark@umich.edu

Ruoxi Jia  
Virginia Tech  
Blacksburg, VA  
ruoxijia@vt.edu

ABSTRACT

It is becoming increasingly common to utilize pre-trained models provided by third parties due to its convenience. At the same time, however, these models may be vulnerable to both poisoning and evasion attacks. We introduce an algorithmic framework that can mitigate potential security vulnerabilities in a pre-trained model when clean data from its training distribution is unavailable to the defender. The framework reverse-engineers samples from a given pre-trained model. The resulting synthetic samples can then be used as a substitute for clean data to perform various defenses. We consider two important attack scenarios—backdoor attacks and evasion attacks—to showcase the utility of synthesized samples. For both attacks, we show that when supplied with our synthetic data, the state-of-the-art defenses perform comparably or sometimes even better than the case when it’s supplied with the same amount of clean data.

1 Introduction

Due to the intensive computation and the substantial amount of data required by training a model from scratch, it is increasingly popular to leverage pre-trained models provided by third parties. Indeed, various online marketplaces, such as BigML and Amazon, have emerged to allow people to buy and sell the pre-trained models. At the same time, recent studies have shown that attackers can manipulate a training dataset to insert backdoor triggers into the learned model [Li et al.(2020a)Li, Wu, Jiang, Li, and Xia] (i.e., backdoor attacks) or manipulate test instances [Szegedy et al.(2013)Szegedy, Zaremba, Sutskever, Bruna, Erhan, Goodfellow, and Fergus] to control the model predictions (i.e., evasion attacks). Thus, it is crucial to strengthen the robustness of pre-trained models against these attacks.

Existing defenses for pre-trained models against both backdoor attacks and evasion attacks rely on the access to a clean dataset drawn from the distribution that the model is trained on. For backdoor attacks, the clean data are needed for both unearthing the potential triggers inserted into the model and further removing the triggers from the model [Zeng et al.(2022)Zeng, Chen, Park, Mao, Jin, and Jia, Wang et al.(2019)Wang, Yao, Shan, Li, Viswanath, Zheng, and Zhao, Chen et al.(2019)Chen, Fu, Zhao, and Koushanfar, Guo et al.(2019)Guo, Wang, Xing, Du, and Song, Liu et al.(2018a)Liu, Dolan-Gavitt, and Garg]. For evasion attacks, the clean data is required to synthesize perturbations that can potentially mislead the model prediction and then fine-tune the model to make it stable to those perturbations [Jeddi et al.(2020)Jeddi, Shafiee, and Wong, Chen et al.(2020)Chen, Liu, Chang, Cheng, Amini, and Wang]. While these defenses have achieved promising results, the requirement of clean in-distribution data could significantly limit their real-world deployment. After all, the main driver of adopting a pre-trained model is in part the lack of in-distribution data.

There have been few attempts in the prior work to lift the requirement of clean in-distribution data. Similar to our idea, DeepInspect [Chen et al.(2019)Chen, Fu, Zhao, and Koushanfar] proposes to synthesize new data samples as a replacement of clean data to perform defenses. Their data synthesis is performed class by class. For each class, one optimizes over the input space of the pre-trained model to seek the image that achieves the lowest loss to be predicted into the class. However, direct optimization over image space often ends up producing meaningless patterns that are far...
Figure 1: Illustration of FrED and the four losses adopted in the clean-data-free case. FrED starts from GAN generated outputs and incorporates two losses from state-of-the-art MI and two novel stability losses to supervise the synthesis: (1) Prior loss $L_P$ measures how realistic a synthesized sample is according to the discriminator; (2) Inward target class loss $L_I$ measures the distance of the synthesized sample to a given class; (3) Look-around stability loss $L_{ARD}$ measures the local stability of the sample; (4) One-step-look-ahead stability loss $L_{AHD}$ measures the synthesized sample’s distance to the target class after one step of virtual tuning.

from the actual clean data distribution. Hence, despite the promising result on simple datasets such as MNIST, this data synthesis approach is limited in offering acceptable defense performance on more complex datasets.

To address the challenge above, we introduce FrED, a new framework for synthesizing samples to support backdoor and evasion defenses for pre-trained models. FrED has the following advantages:

- **State-of-the-art defense performance when clean in-distribution data is unavailable:** On a range of datasets and pre-trained model architectures, employing the synthetic samples produced by FrED can help the state-of-the-art backdoor and evasion defense achieve the performance comparable to or sometimes even better than using the same amount of clean data.

- **Boosting state-of-the-art defense performance when only limited clean data is available.** The state-of-the-art backdoor defense [Zeng et al.(2022)Zeng, Chen, Park, Mao, Jin, and Jia] assumes the availability of limited clean data. FrED is also useful in this scenario. In particularly, FrED can adjust the synthetic samples to the available clean data so that using the combination of both can surpass the defense performance of using the clean data alone.

The design of FrED draws inspiration from the advanced model inversion attack techniques but is substantially novel. In model inversion attacks, the adversary aims at reconstructing the representative input corresponding to a target label based on the knowledge of the model. While such attacks expose privacy about training data and thus are undesirable in most cases, we turn this weakness into a strength by reducing the problem of clean data synthesis for a pre-trained model into that of inverting data from the model. Similar to recent advances of MI attacks [Zhang et al.(2020)Zhang, Jia, Pei, Wang, Li, and Song, Chen et al.(2021)Chen, Kahla, Jia, and Qi], we cope with the high-dimensional search space in image synthesis by optimizing over the latent space of a neural network generator. The generator is trained on a auxiliary dataset that is relevant to the pre-trained model training data. The requirements of the auxiliary dataset are very mild: it may not share classes with the pre-trained models; it could be unlabeled, small, and come from a distribution with significant shifts from the training distribution of the pre-trained model.

Different from existing MI attack techniques, we put an extra emphasis on improving the functionality of the synthetic samples. Specifically, motivated by the fact that pre-trained model reaches convergence when trained with clean data, we design a loss function $L_{ARD}$ to enforce synthetic data to maintain low prediction loss under small perturbations of the pre-trained model parameters. Moreover, given the fact that fine-tuning the pre-trained model with clean data would still achieve good accuracy on themselves, we introduce another loss function $L_{AHD}$ to synthesize data with similar behaviors. Figure 1 illustrates the key ideas of FrED for performing data synthesis. When limited clean in-distribution data are available, we propose to further improve synthetic samples by mimicking the features of clean data.
As shown by extensive evaluation against both backdoor and evasion attacks, the synthetic samples provided by FReD can be as competent as or even better than clean data in terms of maintaining accuracy on clean data and improving robustness of the defenses. One can also leverage the synthetic samples in combination with any available clean data to significantly boost the defense performance. In addition, we validate various design choices of FReD via extensive ablation studies.

2 Related Works

Backdoor defenses. Backdoor defenses normally can be divided into three categories: 1) Poison detection via outlier detection regarding functionalities or artifacts [Gao et al.(2019)Gao, Xu, Wang, Chen, Ranasinghe, and Nepal, Chen et al.(2018)Chen, Carvalho, Baracaldo, Ludwig, Edwards, Lee, Molloy, and Srivastava, Tran et al.(2018)Tran, Li, and Madry, Koh and Liang(2017), Chou et al.(2020)Chou, Tramèr, and Pellegrino, Zeng et al.(2021)Zeng, Park, Mao, and Jia]. 2) Robust training via differential privacy [Du et al.(2019)Du, Jia, and Song, Weber et al.(2020)Weber, Xu, Karlaš, Zhang, and Li], sample quarantine based on training loss [Li et al.(2021)Li, Lyu, Koren, Lyu, Li, and Ma], or ensembled/decoupled pipeline [Levine and Feizi(2020), Jia et al.(2020)Jia, Cao, and Gong, Jia et al.(2021)Jia, Cao, and Gong, Huang et al.(2022)Huang, Li, Wu, Qin, and Ren]. 3) Backdoor removal via trigger synthesis and subsequent unlearning [Wang et al.(2019)Wang, Yao, Shan, Li, Viswanath, Zheng, and Zhao, Chen et al.(2019)Chen, Fu, Zhao, and Koushanfar, Guo et al.(2019)Guo, Wang, Xing, Du, and Song]. The first two lines of works require the access to training data, thus inapplicable to mitigating vulnerabilities of pre-trained models provided by third parties. The last line of works is suitable for pre-trained models but often require a small set of clean data to synthesize triggers and further perform unlearning. A recent work from this line, I-BAU [Zeng et al.(2022)Zeng, Chen, Park, Mao, Jin, and Jia], has achieved the state-of-the-art defense performance against a wide range of existing attacks. The minimum amount of clean samples required by I-BAU was 100 for the CIFAR-10 dataset in the original paper, which has caused a considerable drop in the clean accuracy. In this work, we aim to enable this type of defenses to function effectively without any clean data.

Adversarial Training. To date, the most reliable defenses against evasion attacks (adversarial examples) build upon or are variations of the adversarial training procedure proposed in [Madry et al.(2018)Madry, Makelov, Schmidt, Tsipras, and Vladu]. At a high-level, the approach involves framing training a model against evasion attacks as a minimax game in which during each iteration, the model is trained to overcome the most potent adversarial example at that time. The result is a model that generally has lower accuracy on unaltered data, but has a massive improvement in robustness against evasion attacks. Though the resulting model is not provably robust, adversarial training and its following variations [Uesato et al.(2019)Uesato, Alayrac, Huang, Stanforth, Fawzi, and Kohli, Carmon et al.(2019)Carmon, Raghnathan, Schmidt, Duchi, and Liang, Gong et al.(2021)Gong, Ren, Ye, and Liu, Wong et al.(2020)Wong, Rice, and Kolter] have shown to be robust against various types of attacks. However, all current adversarial training approaches requires access to the same original training data [Jeddi et al.(2020)Jeddi, Shafiee, and Wong, Chen et al.(2020)Chen, Liu, Chang, Cheng, Amini, and Wang, Croce and Hein(2021), Wang et al.(2020)Wang, Chen, Gui, Hu, Liu, and Wang]. In particular, [Jeddi et al.(2020)Jeddi, Shafiee, and Wong] uniquely discusses the application of adversarial training to improve robustness of pre-trained models. Our approach aims to lift the requirement of clean data in adversarial training for pre-trained models - the process which we will call fine-tuning (FT).

Model Inversion. The goal of model inversion (MI) is similar to ours. But from an attack perspective, MI aims to divulge sensitive attributes in the training data, and to achieve this goal, the generated data should have good visual quality. Fredrikson et al. [Fredrikson et al.(2015)Fredrikson, Jha, and Ristenpart] follows the maximum likelihood principle and performs model inversion by searching over the image space for a sample with highest likelihood under the given target model. DeepInspect [Chen et al.(2019)Chen, Fu, Zhao, and Koushanfar] first employs this naive MI to generate a surrogate training set for backdoor unlearning and achieves good results on MNIST and GTSRB. However, during the experiment we find that samples generated by the naive MI approach have bad visual quality and usually fail in downstream defenses on high-dimensional datasets (e.g., PubFig and CIFAR-10). In this paper, we build upon the idea of recent MI works [Zhang et al.(2020)Zhang, Jia, Pei, Wang, Li, and Song, Chen et al.(2021)Chen, Kahla, Jia, and Qi] that searches for an optimal sample in the latent space of a GAN instead of the image space. Even when the GAN is not trained on the in-distribution data of the target model, this idea can greatly help improve the visual quality of synthesized samples. The key innovations that set our work apart from the MI attacks is that we go beyond the traditional “high-likelihood” assumption made in all existing MI works about clean data and further formalize other
plausible assumptions, especially those related to model stability. We show that enforcing the synthetic data to satisfy these assumptions can significantly improve their utility for defenses.

3 Framework Design

Figure 2: The workflow of the FRED framework. FRED consists of three functional procedures to adapt the public knowledge domain to a distribution that best initiates defenses for a given target model: (1) warm-up procedure: adopting a GAN to acquire public knowledge; (2) model-specific synthesis: using different random seeds with the GAN to generate random samples, and then using the target model and the discriminator to adapt the generated sample to the target distribution; and (3) defense step: using the final synthesized samples to initiate clean-data-free defenses.

3.1 Threat Model and Requirements

We follow a standard threat model for mitigating backdoor and evasion attacks for pre-trained models. The defender obtains a target pre-trained model from a third party. For backdoor attacks, the pre-trained model is potentially already embedded with backdoor triggers and in the test phase, the adversary can query the model with samples patched with the specific backdoor triggers to control the model prediction. For evasion attacks, the attacker may query the model with samples perturbed by small, adversarial perturbations to make the model misclassify the samples. The defender aims to mitigate the potential risks with the complete knowledge about the target model in hand but no access or extremely limited access to clean samples that are drawn from the original training distribution of the target model.

4 Evaluation

4.1 Experimental setup

Datasets. We evaluate datasets built for different prediction tasks, including face recognition, traffic sign classification, and general task. For each task, we choose two datasets, one as private data for training the threat model and another as out-of-distribution data (OOD) for training the GAN in the warm-up step. Detailed usage of the datasets is shown in Table 1.

| Datasets     | Face Recognition | Traffic Sign Classification | General Task |
|--------------|------------------|-----------------------------|--------------|
| Private      | PubFig           | UTSR德（Huang et al.2019）| VLAD-TS(Chen et al.2019) |
| OOD          | CelebA           | STSBD                       | STL-10       |

Table 1: Dataset Settings. For each task, we incorporate two datasets, namely the private dataset and the OOD dataset. In clean-data-free cases, the private dataset was only adopted to train the target model. On the other hand, the OOD dataset is adopted as the public knowledge distribution for that specific task.

Baselines. To evaluate FRED, we compare with three baselines: 1) Clean: We adopted the same number of examples as our settings but used the clean samples from the original training set to illustrate the effect with the original data. 2) Out-of-the-distributional (OOD): We use the OOD samples from the public knowledge domain to evaluate the effects directly. 3) naive MI: We adopt the naive MI adopted in [Chen et al.(2019)Chen, Fu, Zhao, and Koushanfar] to demonstrate the limitations from the prior art.
Implementation Details. Before conducting the defense, a batch of samples need to be collected from the step of the model-specific synthesis. During this step, we generate 20 random samples per class for PubFig and GTSRB; 40 random samples per class for CIFAR-10. A detailed study of choosing a proper number of samples to be generated for each class is shown in Section 4.5. We use the same amount of samples for all the methods for a fair comparison. Note that for each setting, we run the defenses for three times and compute the average of the defense performance. As for the hyperparameters used in the model-specific synthesize step, we set $\lambda_1 = 1000$, $\lambda_2 = 1000$, $\lambda_3 = 10$, $\lambda_4 = 1000$ for PubFig and GTSRB, and $\lambda_1 = 1000$, $\lambda_2 = 10$, $\lambda_3 = 1$, $\lambda_4 = 1000$ for CIFAR-10. For the sample generation used for adversarial FT, we evaluate 35 seeds for GTSRB and 200 seeds for CIFAR-10. Unless otherwise noted, all experiments for FT are run for ten instances. The hyperparameter settings are the same.

4.2 FreD Boosted Backdoor Unlearning

To evaluate the functionality and generalization ability of samples generated by FreD, we use I-BAU[Zeng et al.(2022)] which achieves the state-of-the-art performance on backdoor unlearning for a given poisoned model.

Evaluation Metrics. We use two metrics to evaluate the backdoor unlearning performance: 1) ACC: model accuracy on the clean test set; 2) ASR: model attack success rate. Note that we exclude the samples whose true labels are the same as the target poison label when calculating ASR. A well-unlearned model should have a low ASR while maintaining a high ACC.

![Figure 3: Datasets and examples of backdoor attacks incorporated.](image)

Figure 3: Datasets and examples of backdoor attacks incorporated. We consider three different datasets in this work: (1) the CIFAR-10, (2) the GTSRB, and (3) the PubFig. Nine different backdoor attack triggers are included in the experimental part as listed. Above, we also show the target label used during the evaluated attacks (e.g., Hidden targeting at label 8 of the CIFAR-10 dataset).

Attack Settings. For backdoor attacks, we evaluate nine different kinds of backdoor attacks over three different datasets in all-to-one settings (the target model will misclassify all other classes’ samples patched with the trigger as the target class). The details of the adopted trigger’s visual effects over the specific dataset and the target labels are illustrated in Fig. 3. To be more specific, we evaluated the hidden trigger backdoor attack (Hidden) [Saha et al.(2020)], input-aware backdoor (IAB) attack [Nguyen and Tran(2020)], and WaNet [Nguyen and Tran(2021)] on the CIFAR-10. Noticing that initially, the hidden trigger backdoor attack can only work in one-to-one attack settings (only aim to fool one class with the trigger), thus only able to obtain a low ASR in all-to-one settings, we manually enlarged the norm bound to $50/255$ with one round of FT of a pre-trained clean model to achieve an acceptable ASR. We also evaluated $L_0$ invisible ($L_0$ inv) [Li et al.(2020)a] and the frequency invisible smooth (Smooth) attack [Zeng et al.(2021)] on the GTSRB. On the PubFig,

---

1[https://github.com/UMBCvision/Hidden-Trigger-Backdoor-Attacks](https://github.com/UMBCvision/Hidden-Trigger-Backdoor-Attacks)
2[https://github.com/VinAIResearch/input-aware-backdoor-attack-release](https://github.com/VinAIResearch/input-aware-backdoor-attack-release)
3[https://github.com/VinAIResearch/Warping-based_Backdoor_Attack-release](https://github.com/VinAIResearch/Warping-based_Backdoor_Attack-release)
4[https://github.com/YiZeng623/frequency-backdoor](https://github.com/YiZeng623/frequency-backdoor)
we adopted trojan watermark (Troj-WM) [Liu et al.(2018b)Liu, Ma, Aafer, Lee, Zhai, Wang, and Zhang], trojan square (Troj-SQ) [Liu et al.(2018b)Liu, Ma, Aafer, Lee, Zhai, Wang, and Zhang], and blend attack [Chen et al.(2017)Chen, Liu, Li, Lu, and Song]. The details of the implementations of each attack follow the original implementations from the original work.

Figure 4: Examples of images obtained by FRED and naive MI. Each subplot represents randomly generated samples for the same class. The upper row shows image generated by FRED and the lower row shows image generated by naive MI.

Figure 5: Examples of model-specific synthesize by FRED at first 10 iterations for Identity 12 in PubFig dataset. Rightmost in the lower row is the real image of this identity from PubFig.

| Dataset     | Initial Clean | OOD | Naive | FRED | Initial Clean | OOD | Naive | FRED | Initial Clean | OOD | Naive | FRED | Initial Clean | OOD | Naive | FRED |
|-------------|---------------|-----|-------|------|---------------|-----|-------|------|---------------|-----|-------|------|---------------|-----|-------|------|
| PubFig      | 1.00 0.30 0.57 | 0.83 0.07 0.11 | 0.14 0.97 0.21 | 0.03 0.97 0.32 | 1.00 0.30 0.57 | 0.83 0.07 0.11 | 0.14 0.97 0.21 | 0.03 0.97 0.32 | 1.00 0.30 0.57 | 0.83 0.07 0.11 | 0.14 0.97 0.21 | 0.03 0.97 0.32 | 1.00 0.30 0.57 | 0.83 0.07 0.11 | 0.14 0.97 0.21 | 0.03 0.97 0.32 |
| GTSRB       | 0.97 0.98 0.94 | 0.78 0.07 0.21 | 0.05 0.97 0.32 | 0.03 0.97 0.32 | 0.97 0.98 0.94 | 0.78 0.07 0.21 | 0.05 0.97 0.32 | 0.03 0.97 0.32 | 0.97 0.98 0.94 | 0.78 0.07 0.21 | 0.05 0.97 0.32 | 0.03 0.97 0.32 | 0.97 0.98 0.94 | 0.78 0.07 0.21 | 0.05 0.97 0.32 | 0.03 0.97 0.32 |
| CIFAR-10    | 0.94 0.82 0.12 | 0.94 0.82 0.12 | 0.14 0.97 0.21 | 0.03 0.97 0.32 | 0.94 0.82 0.12 | 0.94 0.82 0.12 | 0.14 0.97 0.21 | 0.03 0.97 0.32 | 0.94 0.82 0.12 | 0.94 0.82 0.12 | 0.14 0.97 0.21 | 0.03 0.97 0.32 | 0.94 0.82 0.12 | 0.94 0.82 0.12 | 0.14 0.97 0.21 | 0.03 0.97 0.32 |

Table 2: Results of FRED boosted backdoor unlearning.
Results. To make a fair comparison, we use the same number of images for each class for all the methods. Table 4.2 shows that our method outperforms naive MI and OOD on various datasets and backdoor attacks with higher ACC and lower ASR. One interesting finding is that FRE D achieves a lower ASR and comparable ACC than baseline utilizing clean data when defending against the Hidden attacks performed on the CIFAR-10 dataset. This may be because the model is overfitted to the clean training samples, and samples generated by FRE D reduce the degree of overfitting with more abundant features. FRE D also achieves comparable or slightly lower performance in other situations when compared to the clean baseline. One exception is the Blend attack setting on the PubFig dataset, where even clean data cannot achieve satisfying unlearning performance. This may be because the norm of Blend trigger on the PubFig dataset is too large, making I-BAU hard to find the optimal point during optimization.

The performance improvement achieved by our attack is further corroborated by Figure 4, where we show four randomly generated samples for each dataset. As shown in Figure 4, FRE D tends to generate samples with better visual quality and higher diversity, whereas naive MI generates merely noise-like samples.

To better interpret the process of the model-specific synthesis, we show a series of samples generated at different iterations in Figure 5. We observe that the visual outcomes of generated images varied significantly over the first ten optimization iterations while remaining unchanged to human eyes afterward.

4.3 FRE D Boosted Adversarial Fine-tuning

Evaluation Metrics. As is customary in the adversarial training literature, we evaluate our techniques against two metrics. The first is accuracy on the original, unaltered data (Clean Acc). The second is accuracy under a PGD based attack, which we call robustness. A robustness value of 0% means every adversarial attack is successful. Note that unlike the metrics in the previous section, this value is better when higher and worse when lower. As is customary, we consider PGD with $\epsilon = 8/255$ and $\epsilon = 10/255$ - it is well understood that adversarial training on the PGD attack provides robust defenses against many first-order adversarial attacks. Because our paper aims to illustrate a new technique but not provide a novel defense, we set aside new attacks like AutoAttack [Croce and Hein(2020)] which are designed to mitigate adversarial training.

Attack Settings. We look at models trained on the CIFAR-10 and GTSRB datasets. Our unaltered models - ResNet18 for CIFAR and VGG16 for GTSRB - suffer from 0% accuracy when faced with both types of PGD attacks. We do not consider PubFig because the baseline FT approach on the full original PubFig dataset leads to minimal gains in robustness. In fact, even full end-to-end adversarial training from scratch - which would be considered as the "gold standard" to compare against - leads to relatively minor robustness on this dataset. For the FT algorithm, we borrow ideas from the original paper ([Jeddi et al.(2020)Jeddi, Shafiee, and Wong], specifically gradually increasing the learning rate and then sharply declining. We find that the exact learning rate scheduling proposed in that work does not work for our techniques, so we adjust accordingly.

We compare these approaches with equivalent methods - original data, OOD, and naive - with an equivalent number of samples. We also include results if we were to fine-tune the full training dataset and the full OOD dataset. Fundamentally, there is a trade-off between robustness and clean accuracy - a pattern that has been well-studied in [Tsitpras et al.(2019)Tsitpras, Santurkar, Engstrom, Turner, and Madry, Schmidt et al.(2018)Schmidt, Santurkar, Tsipras, Talwar, and Madry], and so we tailor our experiments as best as possible to have similar clean accuracy values.

Results. The results are shown in Table 4.3. There are a few things to note. Firstly, all techniques (including FT on a small number of the original images) are worse than running FT on the full original training data, which is to be expected. Second, our approach lags slightly behind the original samples’ technique, but it is still comparable. On CIFAR-10, our method achieves 24.69% accuracy under attack with an 83.69% clean accuracy versus 25.45% accuracy under attack with 81.95% clean accuracy. The result on the GTSRB model is even closer: FRE D samples result in a 33.94% robustness accuracy with a 93.66% clean accuracy while using the same amount of the original training data obtains 33.96% robustness with 94.4% clean accuracy. In both cases, fine-tuning on FRE D samples leads to better results than training on samples generated from the naive MI technique, though the result is more apparent on the CIFAR dataset - the naive MI samples result in only a 16.46% robustness gain. Since we imagine the situation where we don’t have access to the original training samples, we also compare against FT using OOD public data. In this case, we show that FRE D surpasses (25.69% for our method versus 16.46% for using OOD on CIFAR) or matches (33.94% vs. 34.23% on GTSRB) training on public data.

One disparity between the results on these two datasets is the effect of FT on the full OOD dataset. In the case of CIFAR-10, FT on the full complementary dataset incurs a significant clean accuracy drop penalty to obtain robustness (clean accuracy of 75.47%) - a penalty that might not be desirable. However, in the case of GTSRB, FT on the full
public dataset is a more viable option, resulting in a drop to 90.92% clean accuracy to obtain 35.06% robustness. However, it bears remembering that in the real world, we would not know a priori how well FT with public data would work with no way of validating its robustness - and we have shown that it is at best uncertain. To note, we do not claim that such an approach is completely impossible and leave robust FT with out-of-distribution datasets for future work to explore. In contrast, adversarial FT with \textsc{Fred} closely matches the results we would obtain using clean samples for both datasets and is a more reliable option.

|               | Initial | Orig. Data | Orig. Data Full | OOD | OOD Full | Naive | FRED |
|---------------|---------|------------|----------------|-----|----------|-------|------|
| \textsc{GTSRB} |         |            |                |     |          |       |      |
| ACC           | 93.79   | 94.4      | 92.54          | 92.3| 90.92    | 93.86 | 93.66|
| PGD (8/255)   | 0.0     | 33.96     | 37.24          | 34.23| 35.06    | 32.81 | 33.84|
| PGD (10/255)  | 0.0     | 28.82     | 33.02          | 30.1| 30.43    | 27.11 | 27.93|
| \textsc{CIFAR-10} |       |           |                |     |          |       |      |
| ACC           | 92.31   | 81.95     | 80.31          | 83.81| 75.47    | 81.5  | 83.69|
| PGD (8/255)   | 0.0     | 25.45     | 46.08          | 16.46| 52.26    | 16.25 | 4.69 |
| PGD (10/255)  | 0.0     | 18.46     | 38.86          | 9.99 | 46.27    | 8.55  | 15.34|

Table 3: Results of \textsc{Fred} based FT on CIFAR-10 and \textsc{GTSRB}. All numbers are accuracies given in %

4.4 Mini Data Booster

In a real scenario, a tiny amount of clean samples might be available. However, the defense performance is usually poor (e.g., high ASR for backdoor unlearning and minimal robustness gain for adversarial FT) when only a small number of clean samples is used. In this section, we show that our proposed \textsc{Fred} can boost the performance of using clean samples only. Specifically, we assume that the defender has a single clean image for each class. Then, we use \textsc{Fred} with the specifically designed feature consistency loss \( L_{\text{CON}} \) to generate 20 additional samples for each class, and the the final result (Booster) is obtained by using the combination of both 1 clean sample and 20 generated samples for each class. Table 4.4 shows that our booster results even outperform the setting in which 20 clean images per class are used for backdoor unlearning.

| PubFig Troj-WM | Clean sample(20 per class) | Clean sample(1 per class) | Booster |
|---------------|-----------------------------|---------------------------|---------|
| ACC           | 0.84                        | 0.44                      | 0.83    |
| ASR           | 0.24                        | 0.35                      | 0       |

| \textsc{GTSRB} \( L_0 \) inv | Clean sample(20 per class) | Clean sample(1 per class) | Booster |
|-----------------------------|-----------------------------|---------------------------|---------|
| ACC           | 0.98                        | 0.95                      | 0.98    |
| ASR           | 0.03                        | 1                         | 0       |

| \textsc{CIFAR-10} IAB | Clean sample(20 per class) | Clean sample(1 per class) | Booster |
|-----------------------|-----------------------------|---------------------------|---------|
| ACC                   | 0.82                        | 0.52                      | 0.83    |
| ASR                   | 0.03                        | 0.18                      | 0.01    |

Table 4: Results of backdoor unlearning performance with a small amount of clean data and generated samples.

|               | Orig. Data (200) | Orig. Data (1) | FRED Booster | FRED Booster only |
|---------------|------------------|----------------|--------------|-------------------|
| ACC           | 81.95            | 85.24          | 83.93        | 83.54             |
| PGD (8/255)   | 25.45            | 9.86           | 26.61        | 26.12             |
| PGD (10/255)  | 18.46            | 4.98           | 18.87        | 18.76             |

Table 5: Results of \textsc{Fred} Boosted FT on CIFAR-10. All numbers are accuracies given in %

When we run boosted adversarial FT on CIFAR-10\(^5\), we also see an improvement in results (Table 4.4). To show that this gain in robustness is not simply because of the including of the original training example, we also run FT with the original sample removed. This suggests that adversarial FT with these samples benefits both from having the original training example but also the increase in quality of the synthesized samples.

4.5 Ablation Study

**Impact of different loss terms.** We proposed two loss terms 1) look-around stability loss \( L_{\text{AHD}} \) and 2) one-step-look-ahead stability loss \( L_{\text{ARD}} \) to improve the utility and functionality of the synthesize samples when the clean

\(^5\) We pick CIFAR to better illustrate the changes in the result
Impact of number of images used for defenses.

We study the performance obtained using various number of seeds for each class on GTSRB dataset. Specifically, we choose the number of seeds for each class to be 1,5,10,15,20,25,30 and evaluate by averaging over 3 runs of the experiments. To better interpret the performance of FrED, we also compared with clean and naive MI baselines using same amount of clean / OOD images. Note that this experiment excludes OOD baseline: we use the target model to generate pseudo-labels for OOD images because the label space of OOD and private domain may not overlap, resulting in the absence or insufficiency of samples for some classes.

Backdoor: Figure 6 shows the results against backdoor attacks. We find that the performance is nearly optimal as long as the number of samples used for each class is above 20. While FrED can maintain a high ACC when larger number of generated samples is used, naive MI suffers a significant ACC drop, indicating the limited utility of naive MI generated samples. We also observe during the experiments that performance of naive MI has large variance when evaluating ACC/ASR over different clean/poison samples. On the other hand, the variance of FrED is similar to the variance of using clean data - indicating good generalization ability of FrED generated samples. Another interesting finding is that when performing unlearning with a small amount of samples (i.e., 1 or 5 samples per class), FrED even achieves higher ACC compared with using clean data.
Evasion: A similar trend is found when using FRED for adversarial FT. FT with a small number of synthesized samples results in less robustness than using the original data of the same amount. However, as we increase the number of seeds, FT with FRED samples mirrors the result of using the original data.

5 Conclusions

References

[Li et al.(2020a)] Yiming Li, Baoyuan Wu, Yong Jiang, Zhifeng Li, and Shu-Tao Xia. Backdoor learning: A survey. arXiv preprint arXiv:2007.08745, 2020a.

[Szegedy et al.(2013)] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. arXiv preprint arXiv:1312.6199, 2013.

[Zeng et al.(2022)] Yi Zeng, Si Chen, Won Park, Zhuoqing Mao, Ming Jin, and Ruoxi Jia. Adversarial unlearning of backdoors via implicit hypergradient. In International Conference on Learning Representations, 2022.

[Wang et al.(2019)] Bolun Wang, Yuanshun Yao, Shawn Shan, Huiyi Li, Bimal Viswanath, Haitao Zheng, and Ben Y Zhao. Neural cleanse: Identifying and mitigating backdoor attacks in neural networks. In 2019 IEEE Symposium on Security and Privacy (SP), pages 707–723. IEEE, 2019.

[Chen et al.(2020)] Si Chen, Mostafa Kahla, Ruoxi Jia, and Guo-Jun Qi. Knowledge-enriched distributional model inversion attacks. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 16178–16187, 2021.
Brandon Tran, Jerry Li, and Aleksander Madry. Spectral signatures in backdoor attacks. In Advances in Neural Information Processing Systems, pages 8000–8010, 2018.

Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In Doina Precup and Yee Whye Teh, editors, Proceedings of the 34th International Conference on Machine Learning, volume 70 of Proceedings of Machine Learning Research, pages 1885–1894, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR. URL http://proceedings.mlr.press/v70/koh17a.html.

Edward Chou, Florian Tramèr, and Giancarlo Pellegrino. Sentinel: Detecting localized universal attacks against deep learning systems. In Deep Learning and Security Workshop, 2020. URL https://arxiv.org/abs/1812.00292.

Yi Zeng, Won Park, Z Morley Mao, and Ruoxi Jia. Rethinking the backdoor attacks’ triggers: A frequency perspective. In ICCV, 2021.

Min Du, Ruoxi Jia, and Dawn Song. Robust anomaly detection and backdoor attack detection via differential privacy. arXiv preprint arXiv:1911.07116, 2019.

Maurice Weber, Xiaojun Xu, Bojan Karlaš, Ce Zhang, and Bo Li. Rab: Provable robustness against backdoor attacks. arXiv preprint arXiv:2003.08904, 2020.

Yige Li, Xixiang Lyu, Nodens Koren, Lingjuan Lyu, Bo Li, and Xingjun Ma. Anti-backdoor learning: Training clean models on poisoned data. Advances in Neural Information Processing Systems, 34, 2021.

Alexander Levine and Soheil Feizi. Deep partition aggregation: Provable defense against general poisoning attacks, 2020.

Jinyuan Jia, Xiaoyu Cao, and Neil Zhenqiang Gong. Intrinsic certified robustness of bagging against data poisoning attacks, 2020. 

Jinyuan Jia, Xiaoyu Cao, and Neil Zhenqiang Gong. Certified robustness of nearest neighbors against data poisoning attacks, 2021.

Kunzhe Huang, Yiming Li, Baoyuan Wu, Zhan Qin, and Kui Ren. Backdoor defense via decoupling the training process. In ICLR, 2022.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In International Conference on Learning Representations, 2018. URL https://openreview.net/forum?id=rJzIBfZAb.

Jonathan Uesato, Jean-Baptiste Alayrac, Po-Sen Huang, Robert Stanforth, Alhussein Fawzi, and Pushmeet Kohli. Are labels required for improving adversarial robustness? In NeurIPS, 2019.

Yair Carmon, Aditi Raghunathan, Ludwig Schmidt, John C Duchi, and Percy S Liang. Unlabeled data improves adversarial robustness. Advances in Neural Information Processing Systems, 32, 2019.

Chengyue Gong, Tongzheng Ren, Mao Ye, and Qiang Liu. Maxup: Lightweight adversarial training with data augmentation improves neural network training. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 2474–2483, June 2021.

Eric Wong, Leslie Rice, and J. Zico Kolter. Fast is better than free: Revisiting adversarial training. ArXiv, abs/2001.03994, 2020.

Francesco Croce and Matthias Hein. Adversarial robustness against multiple lp-threat models at the price of one and how to quickly fine-tune robust models to another threat model. ArXiv, abs/2105.12508, 2021.

Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. Model inversion attacks that exploit confidence information and basic countermeasures. In Proceedings of the 22nd ACM SIGSAC conference on computer and communications security, pages 1322–1333, 2015.
[Pinto et al.(2011)Pinto, Stone, Zickler, and Cox] Nicolas Pinto, Zak Stone, Todd Zickler, and David Cox. Scaling up biologically-inspired computer vision: A case study in unconstrained face recognition on facebook. In CVPR 2011 WORKSHOPS, pages 35–42. IEEE, 2011.

[Houben et al.(2013)Houben, Stallkamp, Salmen, Schlipsing, and Igel] Sebastian Houben, Johannes Stallkamp, Jan Salmen, Marc Schlipsing, and Christian Igel. Detection of traffic signs in real-world images: The German Traffic Sign Detection Benchmark. In International Joint Conference on Neural Networks, number 1288, 2013.

[Krizhevsky et al.(2010)Krizhevsky, Nair, and Hinton] Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. Cifar-10 (canadian institute for advanced research). URL http://www.cs.toronto.edu/~kriz/cifar.html.

[Liu et al.(2015)Liu, Luo, Wang, and Tang] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In Proceedings of the IEEE international conference on computer vision, pages 3730–3738, 2015.

[Huang()LinLin Huang. Chinese traffic sign database. URL http://www.nlpr.ia.ac.cn/pal/trafficdata/recognition.html.

[Coates et al.(2011)Coates, Ng, and Lee] Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In Proceedings of the fourteenth international conference on artificial intelligence and statistics, pages 215–223. JMLR Workshop and Conference Proceedings, 2011.

[Saha et al.(2020)Saha, Subramanya, and Pirsiavash] Aniruddha Saha, Akshayvarun Subramanya, and Hamed Pirsiavash. Hidden trigger backdoor attacks. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 34, pages 11957–11965, 2020.

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

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

[Li et al.(2020b)Li, Xue, Zhao, Zhu, and Zhang] Shao Feng Li, Minhui Xue, Benjamin Zhao, Haojin Zhu, and Xinpeng Zhang. Invisible backdoor attacks on deep neural networks via steganography and regularization. IEEE Transactions on Dependable and Secure Computing, 2020b.

[Li et al.(2018b)Liu, Ma, Aafer, Lee, Zhai, Wang, and Zhang] Yingqi Liu, Shiqing Ma, Yousra Aafer, Wen-Chuan Lee, Juan Zhai, Weihang Wang, and Xiangyu Zhang. Trojaning attack on neural networks. In 25nd Annual Network and Distributed System Security Symposium, NDSS. The Internet Society, 2018b.

[Chen et al.(2017)Chen, Liu, Li, and Song] Xinyun Chen, Chang Liu, Bo Li, Kimberly Lu, and Dawn Song. Targeted backdoor attacks on deep learning systems using data poisoning, 2017.

[Schmidt and Hein(2020)] Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In International conference on machine learning, pages 2206–2216. PMLR, 2020.

[Tsipras et al.(2019)Tsipras, Santurkar, Engstrom, Turner, and Madry] Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness may be at odds with accuracy. In International Conference on Learning Representations, 2019. URL https://openreview.net/forum?id=SyxAb30cY7.

[Schmidt et al.(2018)Schmidt, Santurkar, Tsipras, Talwar, and Madry] Ludwig Schmidt, Shibani Santurkar, Dimitris Tsipras, Kunal Talwar, and Aleksander Madry. Adversarially robust generalization requires more data. In NeurIPS, 2018.