

Don't FREAK Out: A Frequency-Inspired Approach to Detecting Backdoor Poisoned Samples in DNNs

Hasan Abed Al Kader Hammoud¹ Adel Bibi² Philip H.S. Torr² Bernard Ghanem¹

¹ King Abdullah University of Science and Technology (KAUST) ² University of Oxford

Abstract

In this paper we investigate the frequency sensitivity of Deep Neural Networks (DNNs) when presented with clean samples versus poisoned samples. Our analysis shows significant disparities in frequency sensitivity between these two types of samples. Building on these findings, we propose FREAK, a frequency-based poisoned sample detection algorithm that is simple yet effective. Our experimental results demonstrate the efficacy of FREAK not only against frequency backdoor attacks but also against some spatial attacks. Our work is just the first step in leveraging these insights. We believe that our analysis and proposed defense mechanism will provide a foundation for future research and development of backdoor defenses.

1. Introduction

Deep Neural Networks (DNNs) have revolutionized machine learning leading to remarkable advances in various domains such as autonomous vehicles [15], medical imagery analysis [9], and fraud detection [52]. The increased deployment of DNNs in life-critical applications, such as autonomous driving and medical diagnosis, raised concerns particularly with the uncovered vulnerabilities in the form of adversarial attacks.

One extremely insidious form of adversarial attacks is known as backdoor attacks. Backdoor attacks inject malicious behaviour through compromising the training procedure [28, 12], where at inference time, the attacker introduces a special input pattern, known as a trigger, inducing a targeted prediction. The true danger of backdoor attacks lies in their ability to bypass the normal validation procedures that ensure the accuracy and reliability of DNNs [16]. A backdoored model can behave normally on clean inputs and evade detection while misclassifying inputs that contain the trigger leading to severe consequences in high-stakes applications such as action recognition in surveillance systems [19].

Backdoor triggers were typically created in either the

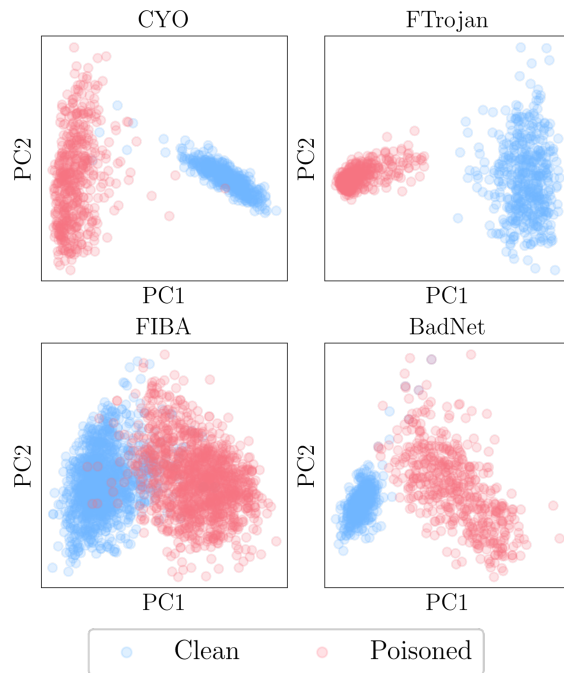


Figure 1: **FREAK PCA Features for Different Attacks.** The 2D PCA projection of the features extracted by FREAK are linearly separable which allows for the successful detection and separation of poisoned and clean samples. This observation holds true for frequency backdoor attacks (CYO, FTrojan and FIBA) and spatial backdoor attacks (BadNet).

spatial [6, 36, 12] or the latent domain [11, 48]. However, recent works have revealed that backdoor attacks could also be created in the frequency domain [18, 13, 47]. Frequency-based backdoor attacks were shown to achieve high attack success rates with a capacity to elude state-of-the-art (SOTA) spatial and latent backdoor defenses. Given that adversaries have the ability to embed their poison in any frequency location across the input image channels, basic filtering techniques such as low-pass, band-pass, or high-pass filtering may not be able to eradicate the trigger.

In response to this challenge, researchers behind frequency-based backdoor attacks have proposed more advanced defenses. For instance, leveraging an autoencoder or JPEG compression [18] to manipulate the Fourier transform of tainted images and filter out the backdoor trigger was shown to be effective. On a different note, FTrojan [47] introduced two adaptive defenses that rely on either anomaly detection or signal smoothing. Nevertheless, these defenses are hampered by one or more limitations: (1) they function in the spatial domain (autoencoder [18]); (2) they can be circumvented by data augmentation (autoencoder and JPEG compression ([18]) and signal smoothing ([47]); (3) they cause significant drops in model accuracy on clean data (signal smoothing ([47]); or (4) they fail to detect the backdoor in the first place (anomaly detection [47]).

In this work, we analyze the distribution of the most sensitive frequency components when the DNN is presented with clean versus poisoned samples. Our analysis reveals that the frequency sensitivity to poisoned samples is considerably distinct from that of clean samples. Drawing on these findings, we present FREAK, a simple yet effective algorithm for identifying poisoned samples based on the distribution of the sensitive frequency components. Our algorithm achieves a high success rate in detecting poisoned samples while maintaining a low false positive rate. Surprisingly, FREAK is not only effective against frequency-based backdoor attacks but also against some spatial backdoor attacks.

2. Related Work

In recent years, a variety of backdoor attacks have been proposed, each of which can differ in two key aspects: the method used to generate the trigger and whether or not the labels are manipulated. In response to these attacks, a number of backdoor defenses have been developed, which can be categorized as follows: (1) defenses aimed at detecting whether a model or a set of samples have been poisoned; (2) defenses aimed at mitigating the backdoor attack; and (3) defenses that aim to both detect and mitigate the attack simultaneously.

Backdoor Attacks. In early backdoor attack methods, backdoor triggers were designed in the spatial domain. For instance, [16] proposed poisoning the data by adding a black square in the corner of a few training samples. [33] solved an optimization problem to find an optimal backdoor trigger for a given mask, such as the square trigger introduced in [16]. However, as research progressed, the importance of invisible triggers that can bypass human inspection became evident, leading to the development of invisible backdoor attacks. This area of research has since evolved, with works such as [34, 6, 27, 7, 46, 12, 1, 51, 38, 30] paving the way. [6] proposed blending the backdoor trigger with clean images instead of stamping it. [27] and [26] adopted

least significant bit and textual string encoding algorithms from steganography to poison the data, respectively. [36] used image warping as a poisoning technique, while [12] emphasized the importance of having learnable transformations to embed an optimal backdoor trigger into the poisoned samples. [7] showed that procedural noise, such as Gabor and Perlin noise, could be used as a backdoor trigger. [44, 2, 45] suggested clean-label backdoor attacks, which apply a backdoor trigger without manipulating the class label of the images.

More recent works suggested exploring alternative domains. For instance, [11] generated imperceptible backdoor triggers by minimizing the Wasserstein distance ([25]) between latent representations of clean and poisoned samples. [50] analyzed the characteristics of spatial backdoor attacks in the Fourier domain and present a technique to generate smooth spatially visible triggers that are smooth in the frequency domain. Finally, and most relevant to our work, [18, 13, 47] showed the power of embedding backdoor attacks in the frequency domain. [18] utilized the concept of Fourier heatmaps from [49] to detect the DNN’s most sensitive frequency bases, which are then used to mount the poisoning information. [13] suggested blending the low-frequency content of a trigger image with those of clean samples to generate poisoned data. [47] converted the color channels from RGB to YUV representation, after which a mix of mid- and high-frequency components is poisoned to bypass possible low-pass or high-pass filtering.

In this work we analyze the properties of frequency-based backdoor attacks. Based on the uncovered properties we propose a new defense.

Backdoor Defenses: As mentioned above, backdoor defenses try to detect the attack [14, 32, 22, 53, 20, 5, 41, 42, 24], mitigate the attack [31, 35, 8, 29], or both detect and mitigate the attack [43, 3, 46, 17, 32, 10, 21, 23, 37, 4].

Early backdoor defenses, such as neural cleanse [46], observed that backdoor attacks create an anomalously small distance between all classes and the poisoned class. On the basis of this observation, the authors proposed solving an optimization problem to detect whether a model has been poisoned after which the backdoor trigger is reverse engineered. Later, improved versions of this defense were introduced by TABOR [17] and ABS [32].

Other backdoor attacks focused on understanding the activations of backdoor attacked models. Fine pruning [31] argued that backdoor attacks could be detected by pruning neurons that are dormant in the presence of clean samples; activation clustering [3] and [43, 21] applied cluster analysis and robust statistics to detect and mitigate backdoor attacks. [10] observed that backdoor attacks shift the network’s attention away from the object, and therefore proposed applying image restoration to reconstruct the spatially poisoned region. Recently, [53] used homology

from topological analysis to uncover structural abnormalities unique to poisoned models.

Unfortunately, existing defenses against frequency backdoor attacks are very scarce and have been shown to fail in certain scenarios. For example, [18] proposed using an autoencoder or JPEG compression to manipulate the Fourier transform of poisoned images and hence neutralize the effect of the backdoor trigger. [47] proposed applying preprocessing techniques such as Gaussian and Wiener filtering to remove the backdoor. However, these defenses cause a huge drop in clean data accuracy or fail to neutralize the backdoor attack.

Considering the limited success of existing defenses in defending against both spatial and frequency backdoor attacks and the critical importance of detecting poisoned samples, FREAK stands out as an effective and necessary addition to the arsenal of defenses against backdoor attacks.

3. Properties of Frequency Backdoor Attacks

Analysis. Existing frequency-based backdoor attacks, e.g., FIBA [13], CYO [18], and FTrojan [47], embed poisoned information into specific frequency bases. Therefore, one would expect that in the presence of the backdoor trigger, such models would attend to the poisoned bases for classifying the poisoned sample. To verify this, we devise an approach similar to the spatial attention technique, GradCAM [40]. The idea is to compute the gradient of the maximal logit of the network with respect to the Fourier transform, specifically, the Fourier magnitude. Afterwards, we visualize the indices of the k most sensitive frequency bases, i.e., top- k gradient values.

Formally, let $f_\theta : \mathcal{X} \rightarrow \mathcal{Y}$ be a classifier parameterized by θ mapping images $x \in \mathcal{X}$ to class labels $y \in \mathcal{Y}$. We denote the most probable class prediction for image x by c_A , where $c_A = \arg \max_{c \in \mathcal{Y}} f_\theta^c(x)$. Let $\mathcal{G}_\eta : \mathcal{X} \rightarrow \mathcal{X}$ denote an attacker-specific poisoned image generator which is parameterized by η . Finally, let $\mathcal{F}(x)$ be the 2D Discrete Fourier Transforms (DFT) of an image x . The gradient we are interested in computing is

$$\nabla_{\text{FREAK}}(x) = \nabla_{|\mathcal{F}(x)|} f_\theta^{c_A}(x). \quad (1)$$

The above quantity expresses the sensitivity of the classifier’s prediction with respect to the Fourier magnitude.

We are interested in comparing the above gradient for clean samples, x_c and their poisoned counterparts $x_p = \mathcal{G}_\eta(x_c)$. Figure 2, presents a binary map that highlights the indices of the top- k values of $\nabla_{\text{FREAK}}(x_c)$ and $\nabla_{\text{FREAK}}(x_p)$ where as Figure 3 shows the distribution of those indices. We make the following key observation; *the frequency bases the network attends to for predicting clean samples*

differ drastically from those for poisoned samples. This observation is the fundamental observation behind FREAK.

FREAK Defense. During inference time, the victim is presented with a sample which may or may not be poisoned. Since the victim has access to clean samples from their test set, we find that a simple mechanism to detect poisoned samples is computing a statistical metric, such as Wasserstein distance, between the distribution of the indices of the top- k most sensitive frequency bases of a sample under inspection sample and that of clean samples. More precisely, we define $\hat{\nabla}_{\text{FREAK}}(x)$ as

$$\hat{\nabla}_{\text{FREAK}}(x)[i, j] = \begin{cases} 1 & \text{if } \nabla_{\text{FREAK}}(x)[i, j] \in \text{top-}k(\nabla_{\text{FREAK}}(x)), \\ 0 & \text{otherwise.} \end{cases} \quad (2)$$

i.e. $\hat{\nabla}_{\text{FREAK}}(x)$ is a binary matrix with value 1 in the locations of the top- k values of $\nabla_{\text{FREAK}}(x)$. The distance between the distribution of the indices of the top- k most sensitive frequency bases can be written as,

$$\gamma(x, y) = d(\text{pool}(x), \text{pool}(y)), \quad (3)$$

where d is the Wasserstein distance and pool is a simple sum-pooling function that aggregates values to obtain a distribution like mapping out of the binary matrices.

The recipe for FREAK is visually presented in Figure 4 and is described below.

1. From the test set \mathcal{D}_{test} , create two subsets of samples, a held-out set $\mathcal{D}_h = \{x_{h_1}, x_{h_2}, \dots, x_{h_n}\}$ and a clean-experimental set $\mathcal{D}_c = \{x_{c_1}, x_{c_2}, \dots, x_{c_m}\}$ where $\mathcal{D}_c \cap \mathcal{D}_h = \emptyset$ and $\mathcal{D}_c \cup \mathcal{D}_h \subseteq \mathcal{D}_{test}$.
2. Compute $\hat{\nabla}_h = \frac{1}{n} \sum_{j=1}^n \hat{\nabla}_{\text{FREAK}}(x_{h_j})$.
3. Compute and store the distance vector $\Gamma_i = \gamma(\hat{\nabla}_{\text{FREAK}}(x_{c_i}), \hat{\nabla}_h)$ for $i = 1, \dots, m$.
4. Fit a Gaussian distribution over the rows of Γ . The obtained distribution is denoted by $\mathcal{N}(\mu_\Gamma, \sigma_\Gamma^2)$ where μ_Γ and σ_Γ^2 are the mean and covariance of the fit Gaussian distribution.
5. Compute the log-likelihood values of the samples of \mathcal{D}_c belonging to the previously fit Gaussian,

$$\mathcal{LL}_{c_i} = \log p(x | \mu_\Gamma, \sigma_\Gamma^2) = -\frac{1}{2} \log(2\pi\sigma_\Gamma^2) - \frac{(x - \mu_\Gamma)^2}{2\sigma_\Gamma^2} \quad (4)$$

for $i = 1, \dots, m$, and store the mean μ_c and the standard deviation σ_c , of the log-likelihood scores.

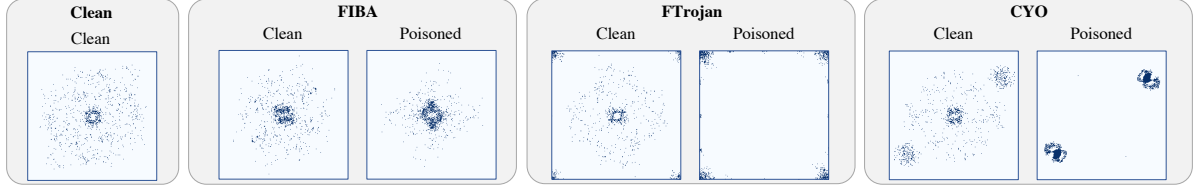


Figure 2: **Visualizing the Indices of the top-k Most Sensitive Frequencies.** By visualizing the top-1000 indices of the FREAK gradient, $\nabla_{\text{FREAK}}(x)$, we can identify the frequency bases that a neural network is most sensitive to for a particular input. We show these indices for a Clean model with a clean input, and for models that have been poisoned with FIBA, FTrojan, and CYO attacks, with both clean and poisoned inputs. This allows us to gain insight into the specific frequencies that are most important to each model and how different attacks affect the network’s sensitivity.

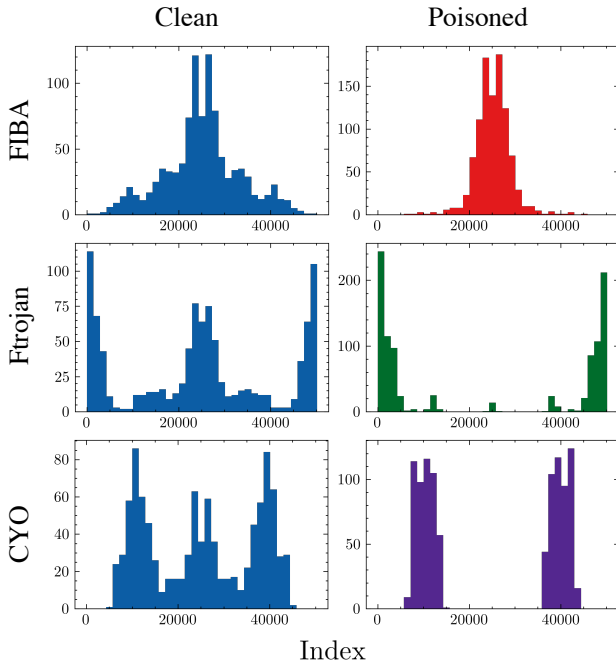


Figure 3: **Visualizing the Distribution of the Indices of the top-k Most Sensitive Frequencies.** Analyzing the distribution of the top-k most sensitive frequency indices can help us detect whether a sample has been poisoned. We compare the distribution shifts for clean and poisoned samples using three attacks: FIBA, FTrojan, and CYO. Backdoored models experience a drastic shift in frequency sensitivity in the presence of the backdoor trigger. This provides valuable insights into the effects of backdoor attacks on the network’s sensitivity to frequency bases.

- When a new sample \tilde{x} is presented for inspection, calculate the distance $\gamma(\tilde{x}, \nabla_h)$ and compute the log-likelihood of the vector belonging to the previously fit Gaussian distribution. If $\mathcal{L}\tilde{x} > \mu_c + \alpha\sigma_c$ or $\mathcal{L}\tilde{x} < \mu_c - \alpha\sigma_c$ then the sample is poisoned, otherwise it’s clean.

4. Experiments

4.1. Experimental Setup & Metrics

Setup. Similar to [18, 12], we conduct our experiments on ImageNet dataset [39]. All models are trained using a ResNet18 trained from scratch using an SGD optimizer with initial learning rate of 0.1 that decays by a factor of 0.25 every 15 epochs. The poisoning rate is fixed to 5.0%.

Backdoor Attack Metrics. To evaluate the performance of the trained backdoor attacked models, we use two commonly used metrics: clean data accuracy (CDA), which measures the DNN’s performance on clean samples, and attack success rate (ASR), which measures the effectiveness of the backdoor attack in instigating the target label. A good backdoored model should have a high ASR and a high CDA.

Detector Metrics. To evaluate the performance of the proposed FREAK detector, we use True Positive Rate (TPR) and False Positive Rate (FPR) as metrics. TPR is a measure of how often a detector correctly identifies a poisoned sample. It is calculated as the number of true positive instances divided by the total number of positive instances. FPR is a measure of how often a detector incorrectly identifies a clean samples as poisoned. It is calculated as the number of false positive results divided by the total number of negative instances. Both TPR and FPR are important metrics in evaluating the performance of a detector. TPR helps us to assess how effective the detector is at identifying poisoned samples, while FPR helps us to identify cases where the detector is misclassifying clean samples as poisoned. A good detector should have a high TPR and a low FPR.

Invisibility Metrics. Following [18], we measure the imperceptibility of an attack using peak signal-to-noise ratio (PSNR) and structural similarity (SSIM). SSIM is a perceptual metric that compares the structural similarity between two images, while PSNR measures the peak signal-to-noise ratio between the two images. A higher SSIM or PSNR value indicates a higher quality image, *i.e. poisoned image looks close to clean one*, while a lower value indicates a more noticeable attack.

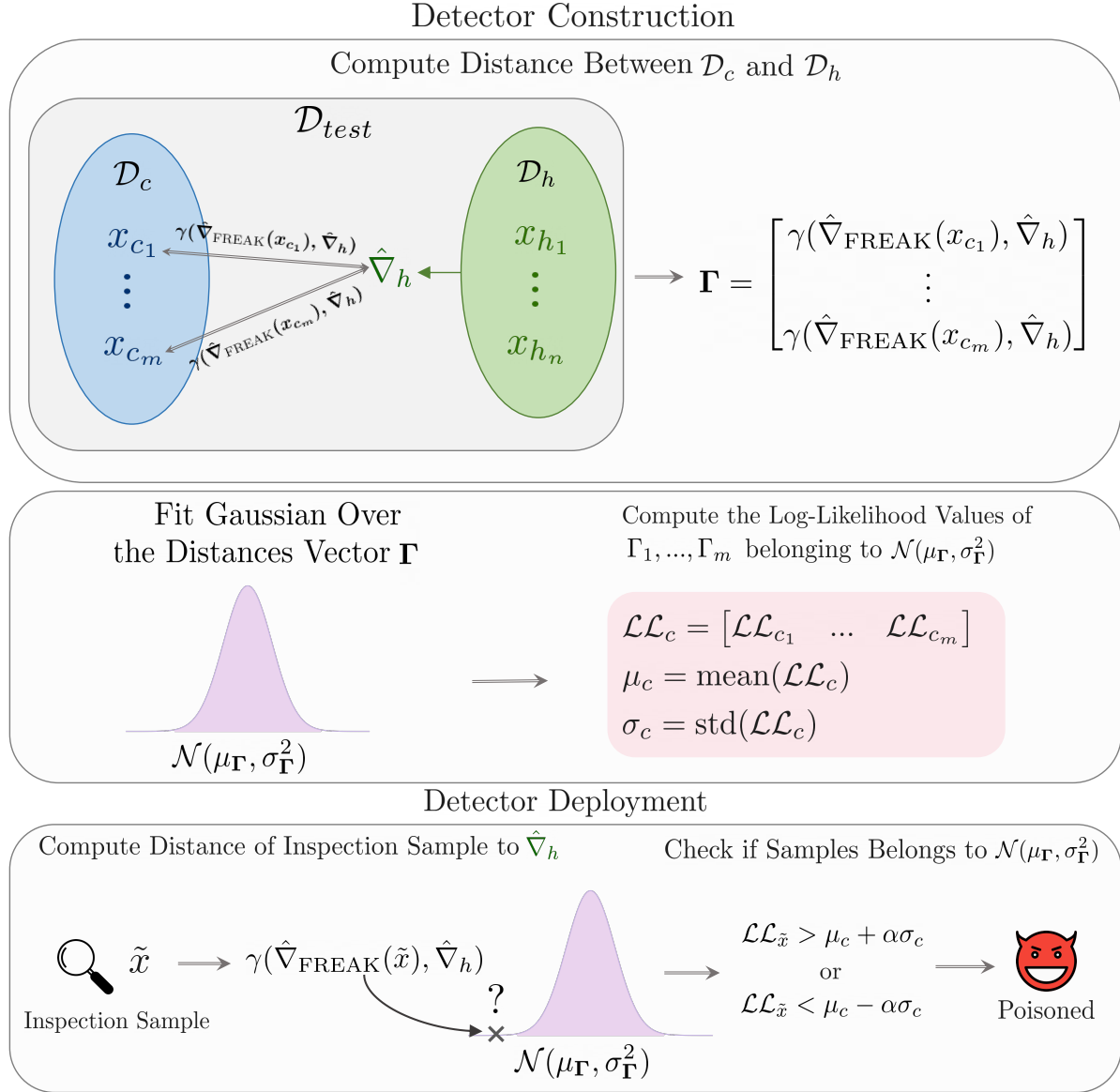


Figure 4: **FREAK Poisoned Sample Detection.** To construct FREAK detector, we first find the distance in Equation 3 between samples from a clean experimental set \mathcal{D}_c and the average distribution of the indices top- k held-out set \mathcal{D}_h , referred to as $\hat{\nabla}_h$. The obtained distances are stored in a vector Γ , whose rows represent the distance of one sample from \mathcal{D}_c to $\hat{\nabla}_h$. Next, we fit a Gaussian distribution over the rows of Γ , referred to as $\mathcal{N}(\mu_\Gamma, \sigma_\Gamma^2)$, and compute the log-likelihood values of the rows of Γ belonging to that distribution. We store the mean and the standard deviation of the log-likelihood values in μ_c and σ_c . When a new sample \tilde{x} is to be inspected, we compute the distance of \tilde{x} to $\hat{\nabla}_h$ and compute the likelihood value of this distance belonging to $\mathcal{N}(\mu_\Gamma, \sigma_\Gamma^2)$, if the value falls within α standard-deviations of the mean then the sample is clean, otherwise it is poisoned.

FIBA [13]		FTrojan [47]			CYO [18]		
h	α	PSNR \uparrow /SSIM \uparrow	Locations	Magnitude	PSNR \uparrow /SSIM \uparrow	k	PSNR \uparrow /SSIM \uparrow
50	0.2	23.98/0.9010	(223,224), (111,111)	30.0	44.89/0.9943	1000	49.51/0.9981

Table 1: **Parameters of Frequency Backdoor Attacks.** The parameters of each frequency backdoor attack are chosen such that an ASR > 95% is achieved. These parameters, along with the invisibility metrics for each attack, are summarized here.

4.2. FREAK against Frequency Backdoor Attacks

We evaluate our backdoor defense against three frequency backdoor attacks, namely, CYO [18], FTrojan [47], and FIBA [13]. As mentioned earlier, the models are trained from scratch using a poisoning rate of 10.0%. Table 1 shows the invisibility metrics (PSNR and SSIM) and hyperparameters selected for each attack. The hyperparameters were chosen to achieve an ASR > 95.0%.

To test our proposed defense, we fix $\alpha = 1$, $|\mathcal{D}_h| = 32$, $|\mathcal{D}_c| = 128$, $\beta = 12$ (pooling filter size), and $k = 5000$. We randomly select 5000 samples from $\mathcal{D}_{\text{test}}$ to be poisoned and another 5000 samples to compute the false positive rate.

FREAK has demonstrated remarkable capabilities in achieving a remarkably high true positive rate (TPR) while simultaneously maintaining an impressively low false positive rate (FPR) in response to all frequency backdoor attacks. Specifically, against CYO [18] and FTrojan [47], FREAK attains a TPR that is close to perfect, i.e., 100%, with an accompanying FPR that is insignificantly close to zero. However, because FIBA [13] corrupts low-frequency data, which typically coincides with the frequencies that a clean network processes, the TPR decreases to 90% with an FPR of 5%. Further details about the results obtained for ResNet34 and ablations of different hyperparameters can be found in the supplementary material.

	TPR (%)	FPR (%)
CYO	99.25	1.56
FTrojan	100.00	1.39
FIBA	90.15	5.31

Table 2: **FREAK Against Frequency Backdoor Attacks.** FREAK proves to be capable of achieving a high TPR while maintaining a low FPR against frequency backdoor-attacks.

4.3. FREAK against Spatial Backdoor Attacks

We subjected FREAK to a series of spatial backdoor attacks, including BadNet [16], Blend [6], SIG [2], and WaNet [36], and summarized the outcomes in Table 3. Interestingly, FREAK was able to achieve a high true positive rate against BadNet and SIG while maintaining a low false positive rate; however, this was not the case for Blend and WaNet, where a significant decline in performance was observed. We discuss this further in the limitations section. The robust performance of FREAK against SIG attacks may be due to the fact that the sinusoidal signal utilized for poisoning the model is of high frequency, creating distinct artifacts in the frequency domain compared to a clean sample.

	TPR (%)	FPR (%)
BadNet	96.60	2.73
SIG	84.51	4.74
WaNet	2.34	1.95
Blend	9.11	3.91

Table 3: **FREAK Against Spatial Backdoor Attacks.** FREAK proves to be capable of achieving a high TPR while maintaining a low FPR on BadNet and SIG, however, this is not the case for WaNet and Blend.

5. Conclusion

Limitations. While our analysis sheds light on the behavior of neural networks when presented with clean and poisoned samples from a frequency-domain standpoint, we acknowledge that the proposed FREAK method is only one possible approach for leveraging these insights, and it may not necessarily be the most optimal. Furthermore, although we opted to use Wasserstein distance, other more suitable distances may be available for this particular problem. Lastly, although our findings encompass all frequency-based backdoor attacks, we acknowledge that we only evaluated a fraction of spatial attacks, and thus further research is required to obtain a more comprehensive assessment.

Conclusion. In conclusion, this paper presents a comprehensive investigation into the frequency sensitivity of Deep Neural Networks when exposed to clean versus poisoned samples. Our results reveal significant differences in frequency sensitivity between these two types of samples. Based on these findings, we propose FREAK, a novel frequency-based poisoned sample detection algorithm that is both simple and effective. Our experimental results demonstrate that FREAK is not only successful against frequency-based backdoor attacks but also some spatial attacks. While our work represents a critical first step towards leveraging these insights, we anticipate that our analysis and proposed defense mechanism will establish a basis for future research and development of backdoor defenses. One possible future direction is sample purification which is presented in the supplementary material.

6. Acknowledgements

This work was supported by the King Abdullah University of Science and Technology (KAUST) Office of Sponsored Research through the Visual Computing Center (VCC) funding, the SDAIA-KAUST Center of Excellence in Data Science and Artificial Intelligence (SDAIA-KAUST AI), and UKRI grant: Turing AI Fellowship EP/W002981/1. We also thank the Royal Academy of Engineering and FiveAI for their support. Adel Bibi has received funding from the Amazon Research Awards.

References

- [1] Eugene Bagdasaryan and Vitaly Shmatikov. Blind backdoors in deep learning models. *ArXiv*, abs/2005.03823, 2021.
- [2] Mauro Barni, Kassem Kallas, and Benedetta Tondi. A new backdoor attack in cnns by training set corruption without label poisoning. *2019 IEEE International Conference on Image Processing (ICIP)*, pages 101–105, 2019.
- [3] Bryant Chen, Wilka Carvalho, Nathalie Baracaldo, Heiko Ludwig, Benjamin Edwards, Taesung Lee, Ian Molloy, and Biplav Srivastava. Detecting backdoor attacks on deep neural networks by activation clustering. In *SafeAI@AAAI*, 2019.
- [4] Huili Chen, Cheng Fu, Jishen Zhao, and Farinaz Koushanfar. Deepinspect: A black-box trojan detection and mitigation framework for deep neural networks. In *IJCAI*, 2019.
- [5] Jian Chen, Xuxin Zhang, Rui Zhang, Chen Wang, and Ling Liu. De-pois: An attack-agnostic defense against data poisoning attacks. *IEEE Transactions on Information Forensics and Security*, 16:3412–3425, 2021.
- [6] Xinyun Chen, Chang Liu, Bo Li, Kimberly Lu, and Dawn Xiaodong Song. Targeted backdoor attacks on deep learning systems using data poisoning. *ArXiv*, abs/1712.05526, 2017.
- [7] Xuan Chen, Yuena Ma, and Shiwei Lu. Use procedural noise to achieve backdoor attack. *IEEE Access*, 9:127204–127216, 2021.
- [8] Hao Cheng, Kaidi Xu, Sijia Liu, Pin-Yu Chen, Pu Zhao, and X. Lin. Defending against backdoor attack on deep neural networks. *ArXiv*, abs/2002.12162, 2020.
- [9] Padideh Danaee, Reza Ghaeini, and David A. Hendrix. A deep learning approach for cancer detection and relevant gene identification. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, 22:219–229, 2017.
- [10] Bao Gia Doan, Ehsan Abbasnejad, and Damith Chinthana Ranasinghe. Februus: Input purification defense against trojan attacks on deep neural network systems. *Annual Computer Security Applications Conference*, 2020.
- [11] Khoa D Doan and Yingjie Lao. Backdoor attack with imperceptible input and latent modification. 2021.
- [12] Khoa D Doan, Yingjie Lao, Weijie Zhao, and Ping Li. Lira: Learnable, imperceptible and robust backdoor attacks. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11946–11956, 2021.
- [13] Yu Feng, Benteng Ma, Jing Zhang, Shanshan Zhao, Yong Xia, and Dacheng Tao. Fiba: Frequency-injection based backdoor attack in medical image analysis. *ArXiv*, abs/2112.01148, 2021.
- [14] Yansong Gao, Chang Xu, Derui Wang, Shiping Chen, Damith Chinthana Ranasinghe, and Surya Nepal. Strip: a defence against trojan attacks on deep neural networks. *Proceedings of the 35th Annual Computer Security Applications Conference*, 2019.
- [15] Sorin Mihai Grigorescu, Bogdan Trasnea, Tiberiu T. Cocias, and Gigel Macesanu. A survey of deep learning techniques for autonomous driving. *Journal of Field Robotics*, 37:362–386, 2020.
- [16] Tianyu Gu, Kang Liu, Brendan Dolan-Gavitt, and Siddharth Garg. Badnets: Evaluating backdooring attacks on deep neural networks. *IEEE Access*, 7:47230–47244, 2019.
- [17] Wenbo Guo, Lun Wang, Xinyu Xing, Min Du, and Dawn Xiaodong Song. Tabor: A highly accurate approach to inspecting and restoring trojan backdoors in ai systems. *ArXiv*, abs/1908.01763, 2019.
- [18] Hasan Abed Al Kader Hammoud and Bernard Ghanem. Check your other door! creating backdoor attacks in the frequency domain. 2021.
- [19] Hasan Abed Al Kader Hammoud, Shuming Liu, Mohammad Alkhrafi, Fahad AlBalawi, and Bernard Ghanem. Look, listen, and attack: Backdoor attacks against video action recognition. *arXiv preprint arXiv:2301.00986*, 2023.
- [20] Zayd Hammoudeh and Daniel Lowd. Simple, attack-agnostic defense against targeted training set attacks using cosine similarity. 2021.
- [21] Jonathan Hayase, Weihao Kong, Raghav Somani, and Sewoong Oh. Spectre: Defending against backdoor attacks using robust statistics. *ArXiv*, abs/2104.11315, 2021.
- [22] Todd P. Huster and Emmanuel Ekwedike. Top: Backdoor detection in neural networks via transferability of perturbation. *ArXiv*, abs/2103.10274, 2021.
- [23] Wei Jiang, Xiangyu Wen, Jinyu Zhan, Xupeng Wang, and Ziwei Song. Interpretability-guided defense against backdoor attacks to deep neural networks. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 2021.
- [24] Kaidi Jin, Tianwei Zhang, Chao Shen, Yufei Chen, Ming Fan, Chenhao Lin, and Ting Liu. A unified framework for analyzing and detecting malicious examples of dnn models. *ArXiv*, abs/2006.14871, 2020.
- [25] Soheil Kolouri, Kimia Nadjahi, Umut Simsekli, Roland Badeau, and Gustavo Kunde Rohde. Generalized sliced wasserstein distances. In *NeurIPS*, 2019.
- [26] Shaofeng Li, Minhui Xue, Benjamin Zi Hao Zhao, Haojin Zhu, and Xinpeng Zhang. Invisible backdoor attacks on deep neural networks via steganography and regularization. *IEEE Transactions on Dependable and Secure Computing*, 18:2088–2105, 2021.
- [27] Yuezun Li, Yiming Li, Baoyuan Wu, Longkang Li, Ran He, and Siwei Lyu. Invisible backdoor attack with sample-specific triggers. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.
- [28] Yiming Li, Baoyuan Wu, Yong Jiang, Zhifeng Li, and Shutao Xia. Backdoor learning: A survey. *ArXiv*, abs/2007.08745, 2020.
- [29] Yiming Li, Tongqing Zhai, Baoyuan Wu, Yong Jiang, Zhifeng Li, and Shutao Xia. Rethinking the trigger of backdoor attack. *ArXiv*, abs/2004.04692, 2020.
- [30] Cong Liao, Haoti Zhong, Anna Cinzia Squicciarini, Sencun Zhu, and David J. Miller. Backdoor embedding in convolutional neural network models via invisible perturbation. *Proceedings of the Tenth ACM Conference on Data and Application Security and Privacy*, 2020.
- [31] Kang Liu, Brendan Dolan-Gavitt, and Siddharth Garg. Fine-pruning: Defending against backdooring attacks on deep neural networks. In *RAID*, 2018.

- [32] Yingqi Liu, Wen-Chuan Lee, Guan hong Tao, Shiqing Ma, Yousra Aafer, and X. Zhang. Abs: Scanning neural networks for back-doors by artificial brain stimulation. *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, 2019.
- [33] Yingqi Liu, Shiqing Ma, Yousra Aafer, Wen-Chuan Lee, Juan Zhai, Weihang Wang, and X. Zhang. Trojaning attack on neural networks. In *NDSS*, 2018.
- [34] Yunfei Liu, Xingjun Ma, James Bailey, and Feng Lu. Reflection backdoor: A natural backdoor attack on deep neural networks. In *ECCV*, 2020.
- [35] Yuntao Liu, Yang Xie, and Ankur Srivastava. Neural trojans. *2017 IEEE International Conference on Computer Design (ICCD)*, pages 45–48, 2017.
- [36] A. Nguyen and A. Tran. Wanet - imperceptible warping-based backdoor attack. *ArXiv*, abs/2102.10369, 2021.
- [37] Ximing Qiao, Yukun Yang, and Hai Helen Li. Defending neural backdoors via generative distribution modeling. In *NeurIPS*, 2019.
- [38] Yankun Ren, Longfei Li, and Jun Zhou. Simtrojan: Stealthy backdoor attack. In *ICIP*, 2021.
- [39] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S. Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115:211–252, 2015.
- [40] Ramprasaath R. Selvaraju, Abhishek Das, Ramakrishna Vedantam, Michael Cogswell, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, 128:336–359, 2019.
- [41] Ezekiel O. Soremekun, Sakshi Udeshi, Sudipta Chattopadhyay, and Andreas Zeller. Exposing backdoors in robust machine learning models. *ArXiv*, abs/2003.00865, 2020.
- [42] Di Tang, Xiaofeng Wang, Haixu Tang, and Kehuan Zhang. Demon in the variant: Statistical analysis of dnns for robust backdoor contamination detection. In *USENIX Security Symposium*, 2021.
- [43] Brandon Tran, Jerry Li, and Aleksander Madry. Spectral signatures in backdoor attacks. In *NeurIPS*, 2018.
- [44] Alexander Turner, Dimitris Tsipras, and Aleksander Madry. Clean-label backdoor attacks. 2018.
- [45] Alexander Turner, Dimitris Tsipras, and Aleksander Madry. Label-consistent backdoor attacks. *ArXiv*, abs/1912.02771, 2019.
- [46] Bolun Wang, Yuanshun Yao, Shawn Shan, Huiying Li, Bimal Viswanath, Haitao Zheng, and Ben Y. Zhao. Neural cleanse: Identifying and mitigating backdoor attacks in neural networks. *2019 IEEE Symposium on Security and Privacy (SP)*, pages 707–723, 2019.
- [47] Tong Wang, Yuan Yao, Feng Xu, Shengwei An, Hanghang Tong, and Ting Wang. Backdoor attack through frequency domain. *ArXiv*, abs/2111.10991, 2021.
- [48] Yuanshun Yao, Huiying Li, Haitao Zheng, and Ben Y. Zhao. Latent backdoor attacks on deep neural networks. *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, 2019.
- [49] Dong Yin, Raphael Gontijo Lopes, Jonathon Shlens, Ekin Dogus Cubuk, and Justin Gilmer. A fourier perspective on model robustness in computer vision. *ArXiv*, abs/1906.08988, 2019.
- [50] Yi Zeng, Won Park, Zhuoqing Morley Mao, and R. Jia. Rethinking the backdoor attacks’ triggers: A frequency perspective. *ArXiv*, abs/2104.03413, 2021.
- [51] J. Zhang, Dongdong Chen, Jing Liao, Qidong Huang, Gang Hua, Weiming Zhang, and Nenghai Yu. Poison ink: Robust and invisible backdoor attack. *ArXiv*, abs/2108.02488, 2021.
- [52] Xinwei Zhang, Yaoci Han, Wei Xu, and Qili Wang. Hoba: A novel feature engineering methodology for credit card fraud detection with a deep learning architecture. *Inf. Sci.*, 557:302–316, 2021.
- [53] Songzhu Zheng, Yikai Zhang, Hubert Wagner, Mayank Goswami, and Chao Chen. Topological detection of trojaned neural networks. *ArXiv*, abs/2106.06469, 2021.