Check Your Other Door! Creating Backdoor Attacks in the Frequency Domain

Hasan Abed Al Kader Hammoud, Bernard Ghanem

King Abdullah University of Science and Technology (KAUST)

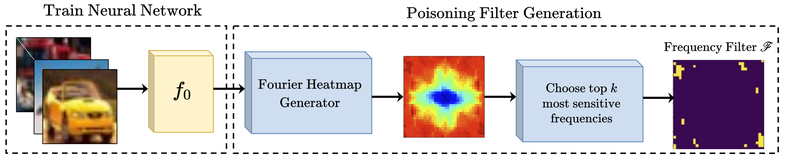
Motivation

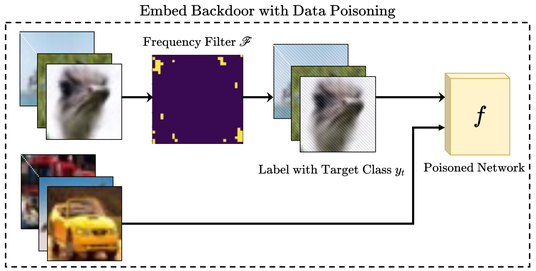
- Current backdoor attacks are limited to either the spatial domain or the latent space domain.
  - ➞ Since existing defenses are based on that prior, we suspect that those defenses would fail in the frequency domain.
  - ➞ Frequency based attacks have been proven to be successful for inference time adversarial attacks, so it would be worth a shot to see if they are also successful against backdoor attacks!

Contributions

- We propose a backdoor attack that utilizes Fourier heatmaps to design a sophisticated backdoor poisoning attack in the frequency domain.
- Unlike previous spatial attacks, our frequency-based attack is completely imperceptible and bypasses spatial defenses.
- We also show two potential ways to defend against frequency-based backdoor attacks and possible ways for the attacker to bypass these defenses.

Our Attack Pipeline





$$\mathcal{M}_{i,j} = \begin{cases} 1 & (i,j) \in \mathbb{I}_k \\ 0 & \text{otherwise} \end{cases} \qquad \mathcal{A}_{\{R,G,B\}_{i,j}} \begin{cases} \neq 0 & (i,j) \in \mathbb{I}_k \\ = 0 & \text{otherwise} \end{cases}$$

$$\mathcal{S}_{\{R,G,B\}} := \mathcal{F}^{-1}\left(\mathcal{F}(\mathcal{S}_{\{R,G,B\}}) \odot (\mathbb{1} - \mathcal{M}) + \mathcal{A}_{\{R,G,B\}}\right)$$

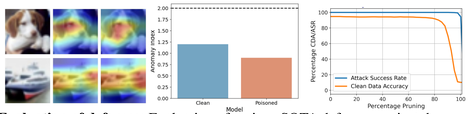
Acknowledgments. This work was supported by the King Abdullah University of Science and Technology (KAUST) Office of Sponsored Research (OSR) under Award No. OSR-CRG2019-4033, as well as the SDAIA-KAUST Center of Excellence in Data Science and Artificial Intelligence (SDAIA-KAUST AI).

Results

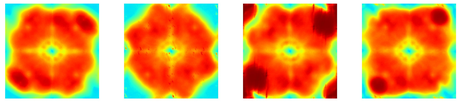
Metric	Ratio	SIG	Refool	SPM	WaNet	FIBA	FTrojan	Poison Ink	Ours
CDA/ASR	3%	89.74 / 99.23	89.20 / 87.16	88.89 / 58.53	91.86 / 32.86	90.92 / 90.10	91.31 / **99.99**	89.65 / 94.22	**92.31** / 99.43
	5%	89.64 / 99.47	89.16 / 89.79	88.90 / 57.69	91.47 / 88.15	90.69 / 95.06	91.64 / 99.10	89.69 / 93.58	**91.88 / 99.88**
	10%	89.45 / 99.40	88.80 / 92.80	89.07 / 57.33	91.22 / 96.96	90.41 / 95.86	90.93 / **100.00**	89.47 / 93.67	**92.10** / 99.97
PSNR↑/LPIPS↓		25.12 / 0.0400	19.38 / 0.0397	38.94 / 0.0001	31.53 / 0.0047	19.40 / 0.0180	41.01 / 0.0001	42.95 / 0.0001	**43.15 / 0.00001**

Comparison between the Proposed Attack and Backdoor Attacks in the Literature. Our frequency-based attack achieves SOTA ASR, CDA, PSNR, and LPIPS metrics. The results shown are for VGG19 trained on CIFAR10. Legend: **First Best**, Second Best

	Poisoning Rate	CDA(%)	ASR(%)
CIFAR10	0.0%	93.92	-
	0.1%	94.00	1.54
	0.2%	94.14	72.31
	0.4%	94.20	85.05
	1.0%	94.38	99.44
	3.0%	94.31	99.79
CIFAR100	0.0%	75.95	-
	0.1%	75.76	60.57
	0.2%	75.75	92.78
	0.4%	75.92	96.49
	1.0%	76.05	98.99
	3.0%	75.36	99.93
GTSRB	0.0%	97.11	-
	0.1%	97.09	71.12
	0.2%	97.19	89.59
	0.4%	97.33	98.04
	1.0%	97.25	98.62
	3.0%	97.47	99.80
ImageNet	0.0%	67.51	-
	0.5%	67.38	0.17
	1.0%	67.13	87.74
	2.0%	67.26	98.01
	3.0%	67.26	98.32



Evaluation of defenses: Evaluation of various SOTA defenses against the proposed frequency-based attack shows the power of the proposed method in evading the defenses. **(a) Grad-CAM** shows high similarity in the attention regions for poisoned and nonpoisoned models; **(b) Pruning** the poisoned model maintains high ASR even after significant drop in CDA. **(c) Neural Cleanse** anomaly indices fall below the anomaly threshold (2.0).



Heatmaps of Various Frequency Selection: Fourier heatmaps of frequency based backdoor attacks with different frequency selection schemes: (a) Clean Model; (b) Random Frequency Selection; (c) Bottom-k Frequency Selection; (d) Proposed Top-k Frequency Selection

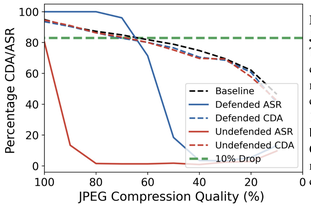
Evaluation of the proposed backdoor attack. We benchmark using ResNet18 trained on various datasets and poisoning rates. Our attack maintains CDA, while achieving high ASR even with small poisoning rates.

	Poisoning Rate	1%	2%
Random (1)	CDA(%)/ASR(%)	67.24/53.91	66.83/60.49
Random (2)		67.23/56.88	66.80/66.11
Bottom-k (1)		67.03/22.58	66.80/55.96
Bottom-k (2)		67.04/0.31	67.02/92.81
Top-k (1)		67.13/87.74	67.26/98.01

Effect of Different Frequency Selection Schemes: Results for frequency filters generated using least sensitive, most sensitive and random frequencies. Choosing the top-k most sensitive frequencies provides the highest ASR among those options.

Method	PSNR↑	SSIM↑	LPIPS↓
BadNets [16]	27.03	0.9921	0.0149
Blend [5]	19.18	0.7291	0.2097
SIG [1]	25.12	0.8988	0.0532
Refool [35]	16.59	0.7701	0.2461
SPM [30]	38.65	0.9665	0.0022
Poison Ink [56]	41.62	0.9915	0.0020
FTrojan [48]	44.87	0.9942	0.0005
FIBA [11]	18.05	0.8077	0.1113
Ours (ResNet18)	47.26	**0.9998**	0.0006
Ours (ResNet34)	47.55	**0.9998**	0.0004
Ours (ResNet50)	46.90	**0.9998**	0.0009
Ours (DenseNet121)	47.21	**0.9998**	**0.0001**
Ours (VGG19)	46.19	**0.9998**	0.0008

Comparing Invisibility Metrics of Backdoor Attacks on ImageNet. Our attack achieves the best invisibility scores compared to other existing methods.



Defending with JPEG Augmentation. Training on JPEG compressed images maintains a high ASR even after a drop of 10% in CDA. The baseline denotes the CDA of the baseline model evaluated on 0 compressed images.