F-mixup: Attack CNNs From Fourier Perspective

Xiu-chuan Li Xu-yao Zhang Fei Yin Cheng-Lin Liu

National Laboratory of Pattern Recognition
Institute of Automation, Chinese Academy of Sciences

Dec 8, 2020

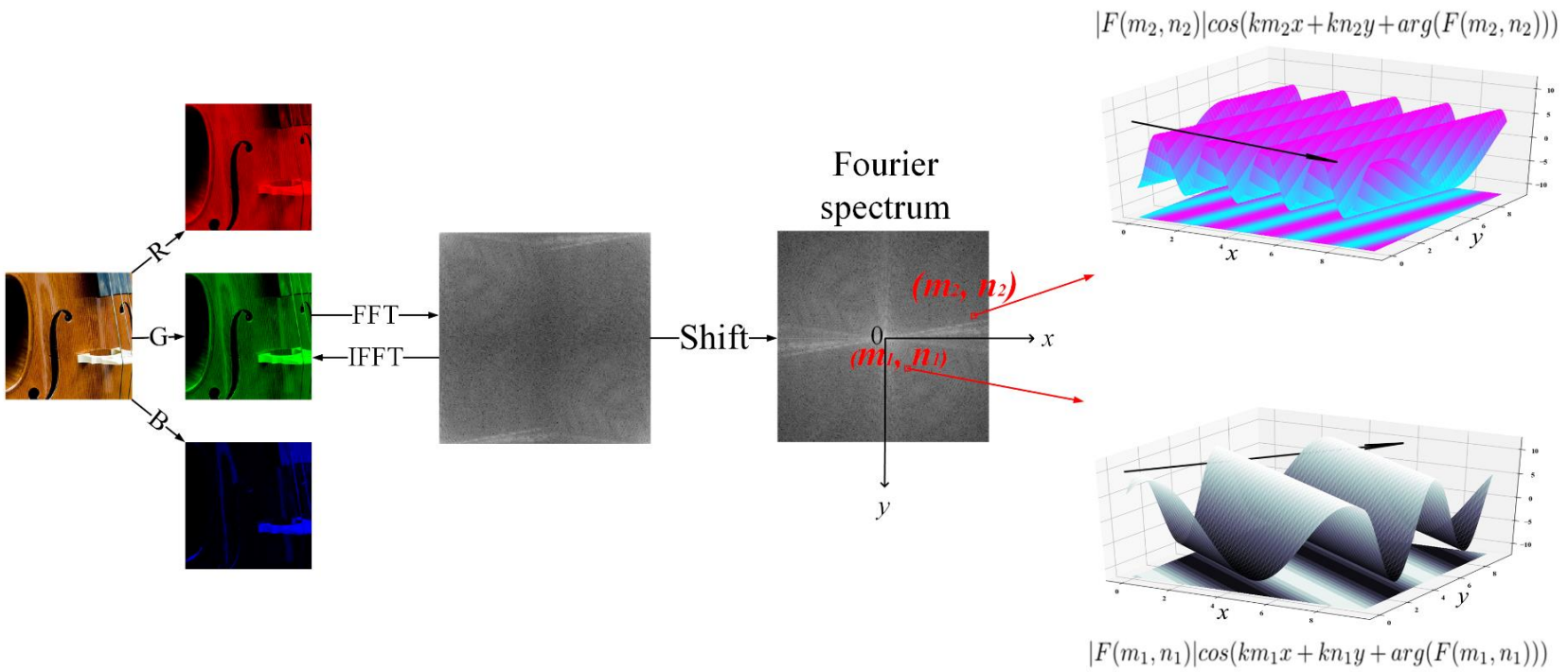
Outline

- Image In frequency domain
- F-mixup
- Experiments
- Conclusion

Image in frequency domain

2D FFT: $$F(u,v) = \frac{1}{M^2}\sum_{x=0}^{M-1}\sum_{y=0}^{M-1} f(x,y)e^{-2\pi\frac{ux+vy}{M}j}$$

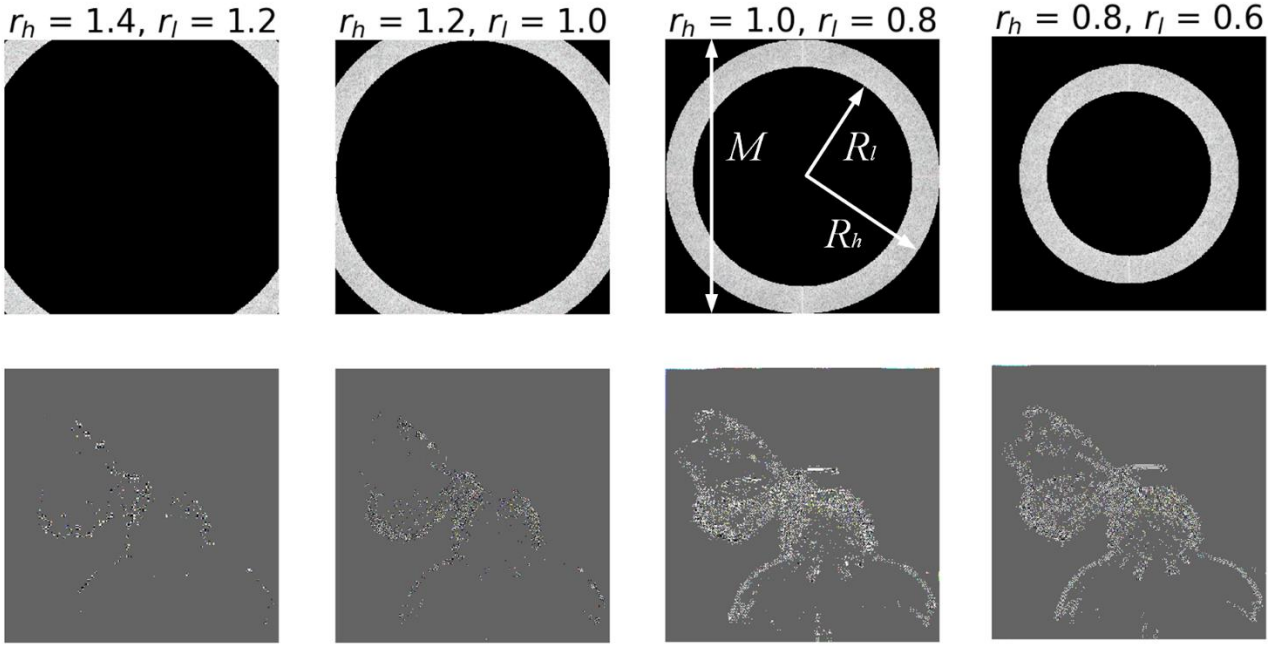
2D IFFT: $$f(x,y) = \sum_{u=0}^{M-1}\sum_{v=0}^{M-1} F(u,v)e^{2\pi\frac{ux+vy}{M}j}$$



Property

(1) Natural images have the bulk of their energy concentrated on the low frequency domain.

(2) High frequencies contain features that are highly predictive, although they are slight thus imperceptible to HVS (human visual system).



F-mixup

Problem Definition

Targeted model: $C(\cdot)$

Clean example: $\boldsymbol{x}$ labeled y, $C(\boldsymbol{x}) = y$

Adversarial attack:

Generate an example $\tilde{\boldsymbol{x}}$ such that $C(\tilde{\boldsymbol{x}}) \neq y$ and $d(x, \tilde{x}) < \rho$, where $d(\cdot,\cdot)$ is a distance metric such as L_2 distance.

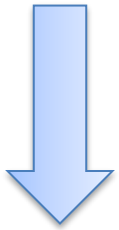
Black-box:

Attacker can only make queries to probe the top-1 label.

Perceptual disparities

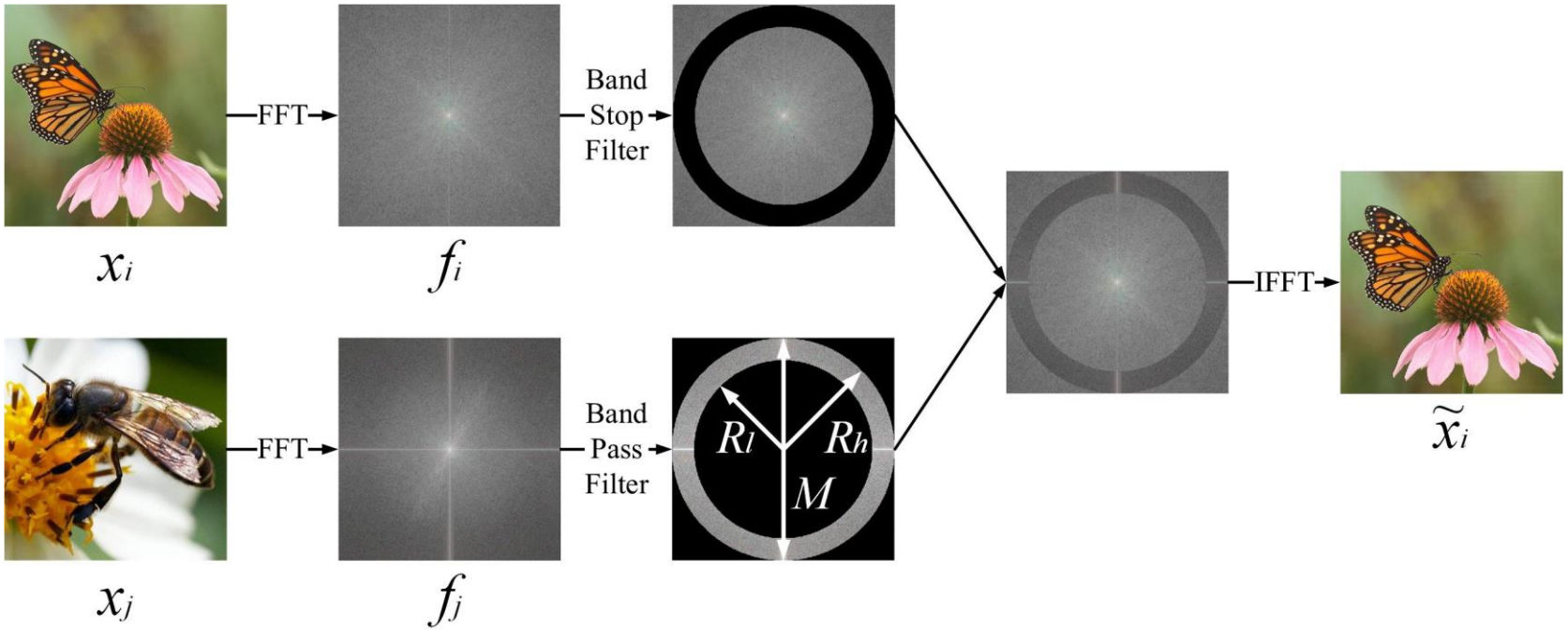
(1) HVS is insensitive to high frequencies.

(2) CNN can exploit the high-frequency image components that are imperceptible to HVS.



Perturbations in high frequency domain may cause CNN make wrong predictions while not be recognized by HVS.

F-mixup



$$\tilde{\boldsymbol{x}}_i = bsf(\boldsymbol{x}_i; r_l, r_h) + bpf(\boldsymbol{x}_j; r_l, r_h)$$

$$r_h = \frac{2R_h}{M}, r_l = \frac{2R_l}{M}$$

Experiments

Setup

Dataset: c

Targeted r

comparisc

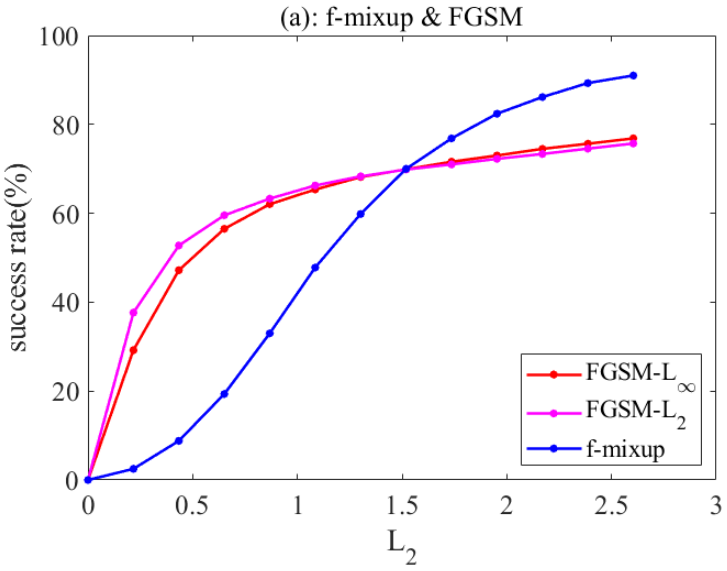
(1) White

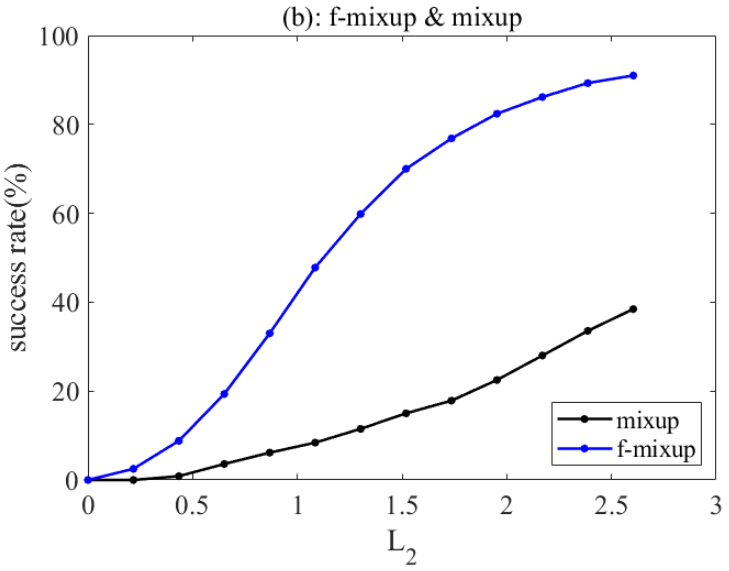
(2) Black-

(3) Gray-b

Comparison with FGSM and mixup

- **FGSM**: gradient information guide attacker in searching adversarial examples
- **mixup:** $\tilde{\boldsymbol{x}}_i = \lambda \boldsymbol{x}_i + (1-\lambda)\boldsymbol{x}_j$





Comparison with QL and SimBA

- **QL**: employs projected gradient descent (white-box attacks) with the estimated gradient to construct adversarial examples.
- **SimBA**: picks one color of a single randomly chosen pixel and perturbs image by increase or decrease the color iteratively.

	Queries	Success Rate	Average L_2 norm
QL	1000	86%	1.319
	500	78.4%	1.393
	200	67.7%	1.637
SimBA	1000	92.7%	1.163
	500	87.5%	1.46
	200	70.4%	1.608
f-mixup	1000	82.1%	1.532
	500	78.2%	1.536
	200	**73.1%**	**1.552**

Conclusion

- Propose a novel black-box attack f-mixup in frequency domain.
- Reveal the preference to high frequencies of CNN compared to HVS.
- Future work:
  - Extensive comparison with more algorithms.
  - Exploration of the different working mechanism between CNN and HVS.

References

1. D. Yin, R. Gontijo Lopes, J. Shlens, E. D. Cubuk, and J. Gilmer. A fourier perspective on model robustness in computer vision. *NIPS*, 2019.

2. H. Wang, X.Wu, Z.Huang, and E.P. Xing. High-frequency component helps explain the generalization of convolutional neural networks. *CVPR*, 2020.

3. A. Ilyas, L. Engstrom, A. Athalye, and J. Lin. Black-box adversarial attacks with limited queries and information. *ICML*, 2018.

4. C. Guo, J.Gardner, Y. You, A. G. Wilson, and K. Weinberger. Simple black-box adversarial attacks. *ICML*, 2019

Thank you!