F-mixup: Attack CNNs From Fourier Spectrum

Xiu-Chuan Li Xu-Yao Zhang Fei Yin Cheng-Lin Liu

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

Introduction

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:

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

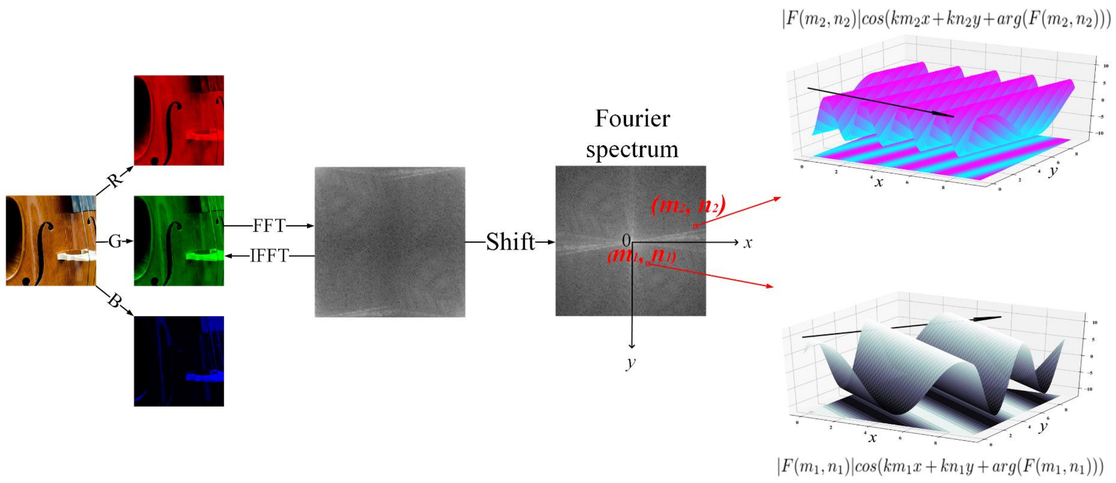

Figure 1: 2D FFT decomposes a 2D function into a group of sinusoidal plane-waves

Property 1

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

Property 2

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

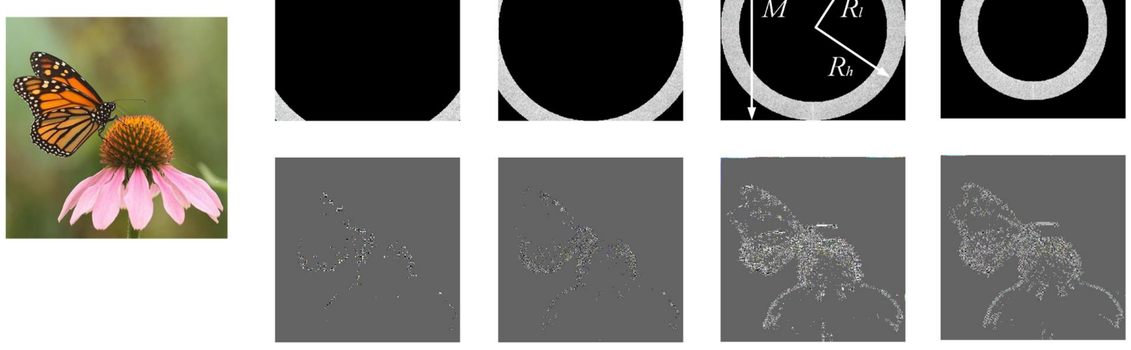

Figure 2: Feature information in high frequency domain

Problem Definition

Targeted model: $C(\cdot)$

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

Adversarial attack:

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

Black-box:

Attacker can only make queries to probe top-1 label

Basic Premise

Premise 1

HVS is insensitive to high frequencies

Premise 2

CNN can exploit the high frequency image components that are imperceptible to HVS.

F-mixup

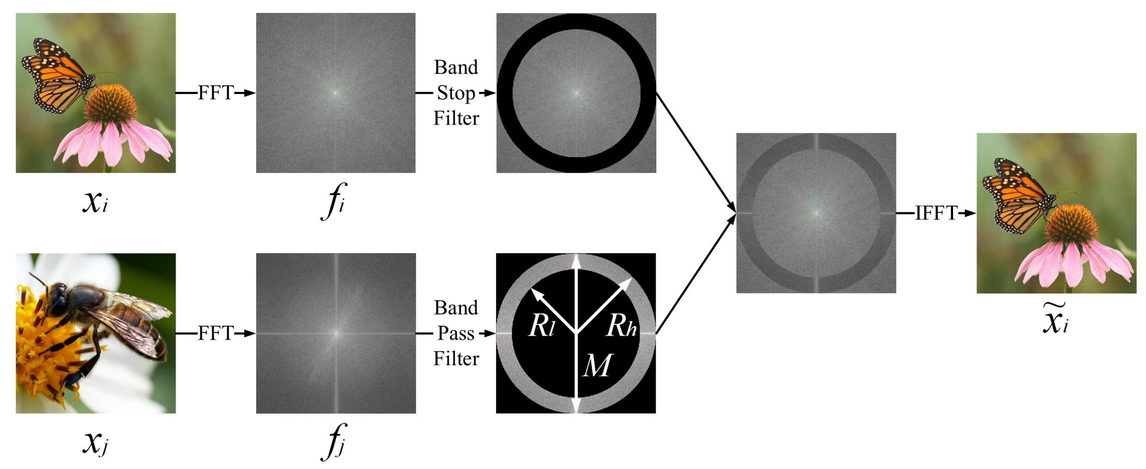

Figure 3: f-mixup between x_i and x_j

Adversarial example:

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

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

Experiments

Setup

- Dataset: cifar10
- Targeted model: VGG16

Comparison with FGSM & mixup

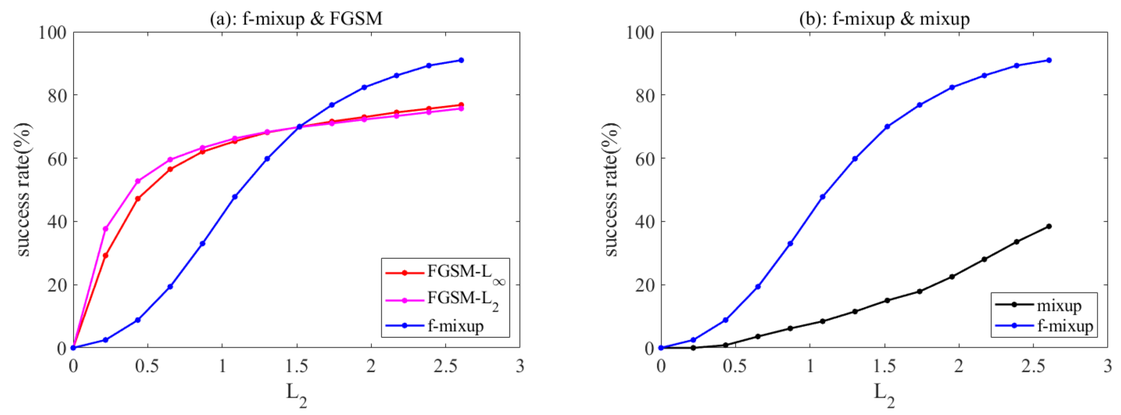

Figure 4: success rate w.r.t. L_2

Comprison with QL & SimBA

	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**