BlurNet: Defense by Filtering the Feature Maps

Ravi Raju

September 26, 2019

WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON

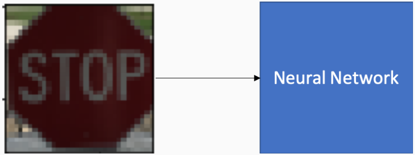
PHARM

Outline

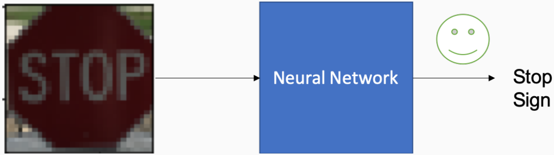
Introduction

Introduction

Vulnerabilities in NNs



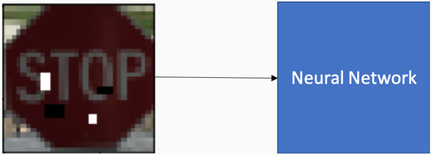
Vulnerabilities in NNs



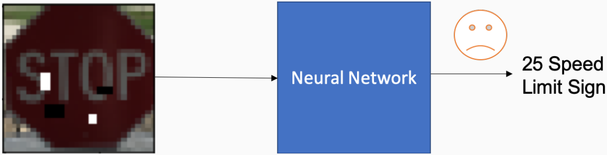
Neural Network

Stop Sign

Vulnerabilities in NNs



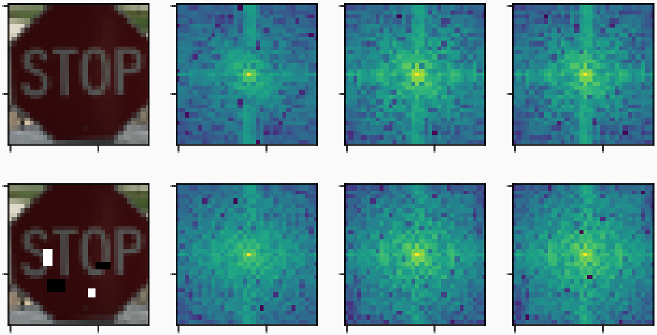
Vulnerabilities in NNs



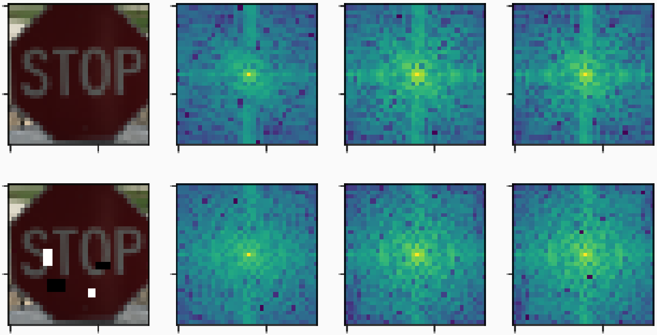
Neural Network

25 Speed Limit Sign

FFT Spectrum of channels

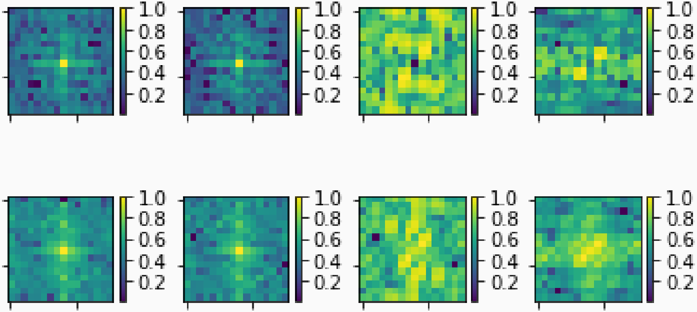


FFT Spectrum of channels



- Log-shifted and normalized frequency spectrum of RGB channels of a natural and perturbed stop sign image
- Lower freqencies correspond to the center and higher ones to the edge.

FFT of First Layer

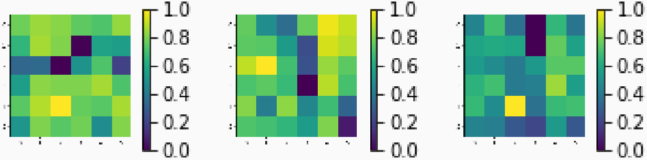


Filtering input vs. Filtering Feature Maps

Table 1: Results from black box evaluation

	Accuracy	**Attack Success Rate**
Baseline	100%	90%
Input filter 3x3	100%	87.5%
Input filter 5x5	100%	67.5%
3x3 filter on L1 feature maps	100%	65%
5x5 filter on L1 feature maps	87.5%	17.5%

Filtering in the higher layers

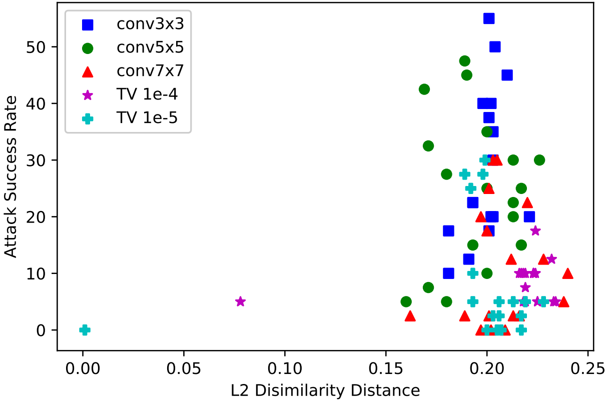


Whitebox Evaluation

Table 2: Results from white box evaluation

	α	**Legitimate Acc.**	**Average Success Rate**	**Worst Success Rate**	L_2 **Distortion**
Baseline	0	91%	49.18%	90%	0.207
3x3 conv	10^{-5}	86.3%	30%	55%	0.201
5x5 conv	0.1	86.3%	24.11%	47.5%	0.189
7x7 conv	0.1	87%	11.61%	30%	0.203
TV	10^{-4}	85.6%	7.92%	17.5%	0.224
TV	10^{-5}	82.3%	8.47%	30%	0.199

L2 vs Attack Plot


conv3x3
conv5x5
conv7x7
TV 1e-4
TV 1e-5
Attack Success Rate
L2 Disimilarity Distance