Defense Mechanism Against Adversarial Attacks Using Density-based Representation of Images

Yen-Ting Huang, Wen-Hung Liao, Chen-Wei Huang

Dept. of Computer Science, National Chengchi University, Taipei, TAIWAN
Pervasive Artificial Intelligence Research (PAIR) Labs, TAIWAN
Email : {ythuang, whliao}@nccu.edu.tw

Abstract

Adversarial examples are slightly modified inputs devised to cause erroneous inference of deep learning models. Protection against the intervention of adversarial examples is a fundamental issue that needs to be addressed before the wide adoption of deep-learning based intelligent systems. In this research, we utilize the method known as input recharacterization to effectively eliminate the perturbations found in the adversarial examples. By converting images from the intensity domain into density-based representation using halftoning operation, performance of the classifier can be properly maintained. With adversarial attacks generated using FGSM, I-FGSM, and PGD, the top-5 accuracy of the hybrid model can still achieve 80.97%, 78.77%, 81.56%, respectively. Although the accuracy has been slightly affected, the influence of adversarial examples is significantly discounted. The average improvement over existing input transform defense mechanisms is approximately 10%.

Methodology

- **Change of Decision Boundaries for input recharacterization**

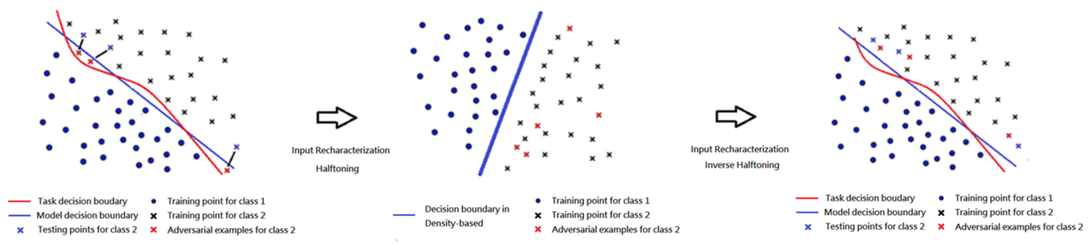

Figure 1: Possible means of recharacterizing adversarial input.

- **Verification of input recharacterization**

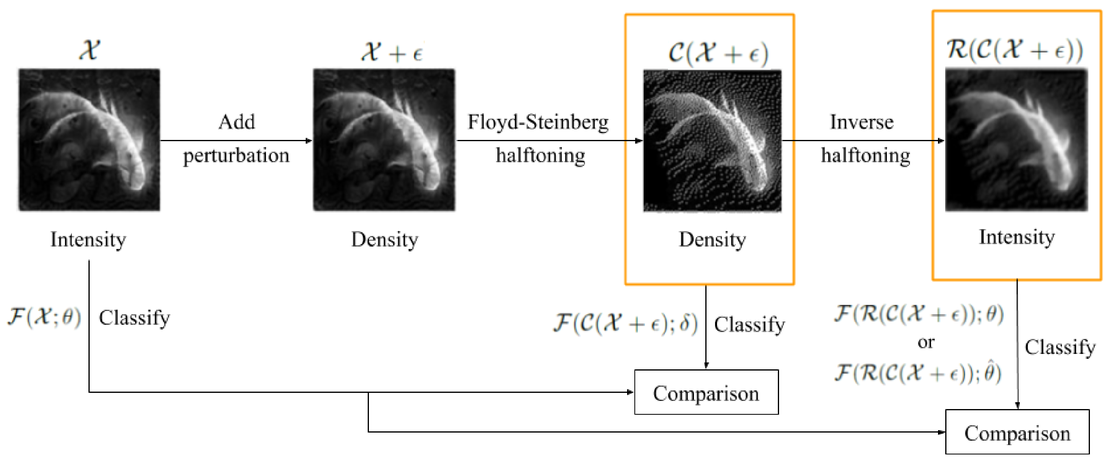

Figure 2: Flowchart of our defense mechanism.

Input recharacterization can consist of two stages: a forward conversion ($\mathcal{C}$) and an optional backward reconstruction ($\mathcal{R}$). We would like to verify if one of the following three conditions will be satisfied using the proposed transformation.

$$\mathcal{F}(\mathcal{C}(\mathcal{X}+\epsilon);\delta) = \mathcal{F}(\mathcal{X};\theta) \tag{1}$$

$$\mathcal{F}(\mathcal{R}(\mathcal{C}(\mathcal{X}+\epsilon));\theta) = \mathcal{F}(\mathcal{X};\theta) \tag{2}$$

$$\mathcal{F}(\mathcal{R}(\mathcal{C}(\mathcal{X}+\epsilon));\hat{\theta}) = \mathcal{F}(\mathcal{X};\theta) \tag{3}$$

Experimental Results

- **Transferability of Adversarial Examples**

Attack	Accuracy	Cropping & Rescaling	TVM	Grayscale	Halftone	Hybrid (intensity)	Hybrid (density)
Baseline	Top-1	56.98	59.13	62.0	61.1	66.01	60.06
	Top-5	77.23	78.56	76.5	80.4	85.14	82.31
FGSM	Top-1	43.65	36.46	12.0	57.78	59.93	59.40
	Top-5	69.96	69.07	31.4	80.34	81.13	80.97
I-FGSM	Top-1	45.10	43.15	10.1	52.01	34.93	52.51
	Top-5	72.52	70.21	17.4	78.35	69.31	78.77
PGD	Top-1	45.68	39.13	10.1	57.23	48.69	58.03
	Top-5	73.26	67.29	17.4	80.91	77.46	81.56

Table 1:Performance of different input transform schemes

- **Launching Attacks in the Halftone Domain**
  - **Global Adversarial Perturbations: PGD Attack**

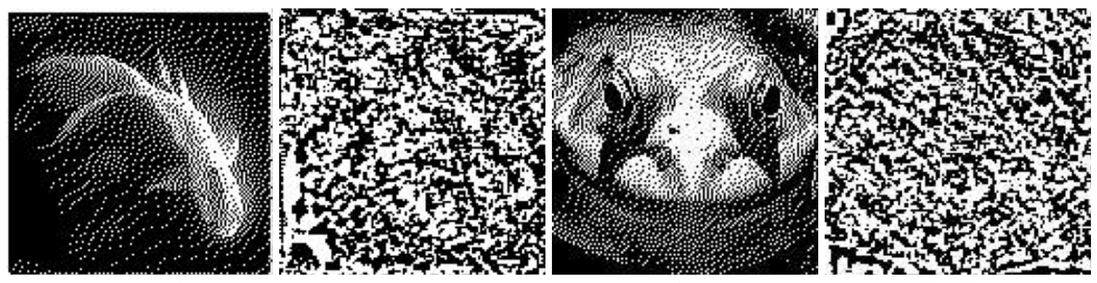

Figure 3:Adding global perturbations in the halftone domain using PGD.

  - **Local Adversarial Perturbations: JSMA Attack**

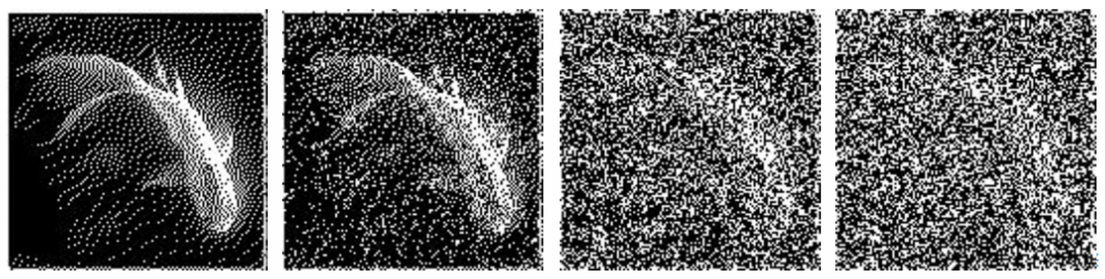

Figure 4:Generating different levels of local perturbations using JSMA.

- **Feasibility of Invalidating Attacks with Two-stage Input Recharacterization**

Attack	Accuracy Defense	Grayscale (Original)	Grayscale (Inverse)	Hybrid (Original)	Hybrid (Inverse)
Baseline	Top-1	62.0	12.0	66.01	26.32
	Top-5	76.5	27.9	85.14	46.64
FGSM	Top-1	12.0	9.8	59.93	23.11
	Top-5	31.4	24.1	81.13	42.26
I-FGSM	Top-1	10.1	8.30	34.93	20.63
	Top-5	17.4	22.05	69.31	40.23
PGD	Top-1	10.1	9.33	48.69	21.57
	Top-5	17.4	23.41	77.46	41.50

Table 2:One-way vs. two-stage transformation for defending adversarial attacks

Acknowledgements

This work was partially supported by The Ministry of Science and Technology, Taiwan, under GRANT No. MOST108-2221-E-004-008 and MOST109-2634-F-004-001 through Pervasive Artificial Intelligence Research (PAIR) Labs.