Project Presentation for
CS 6968: Machine Learning and Optimization

Robust Adversarial ℓ_p Attack Detection Using Block-Sparse Decomposition

Group - **G4**

Emre Acarturk[1]

Maruf A. Mridul[2]

Thomas M. Waite[2]

Spring 2023

[1] Department of Electrical and Computer Systems Engineering
Rensselaer Polytechnic Institute
Tory, NY 12180, USA
acarte@rpi.edu

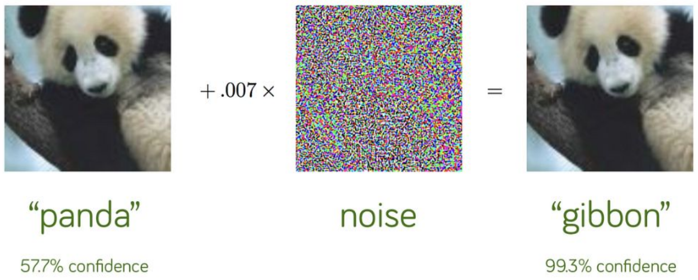
[2] Department of Computer Science
Rensselaer Polytechnic Institute
Tory, NY 12180, USA
{mridum, waitet}@rpi.edu

Overview

- Introduction – Background, Setting and Our Contributions
- Summary of Inspiration Paper (Thaker et al., 2022)
  - Formulation
  - Objective and Algorithm
  - Results
- Our Proposed Approaches
  - Light comparison w/ prev paper
  - Equations, Algorithms
  - Experiments and Results
- Conclusion and Future Works

Introduction

- Adversarial attack



- Different adversarial attack and corresponding defense schemes have been proposed in recent years (Madry et al., 2017, Athalye et al., 2018 etc).
- There is a recent paper (Thaker et al., 2022) on attack classification for block sparse ℓ_p norm bounded attacks.

Our Contributions

- We reproduce their results to validate their claims
- Using the setting and ideas from (Thaker et al., 2022), we propose the following:
    - Both an attack detection and identification algorithm as well as a robust detection algorithm

Inspiration Paper

Darshan Thaker, Paris Giampouras, and René Vidal. Reverse engineering ℓp attacks: A block-sparse optimization approach with recovery guarantees. In *International Conference on Machine Learning*, pp. 21253–21271. PMLR, 2022.

Formulation – Block Sparsity

- Block sparse formulation

$$\min_{\mathbf{c}_s,\mathbf{c}_a} \sum_{i=1}^{r} \|\mathbf{c}_s[i]\|_2 + \sum_{i=1}^{r}\sum_{j=1}^{a} \|\mathbf{c}_a[i][j]\|_2 \quad \text{s.t.} \quad \mathbf{x}' = \mathbf{D}_s\mathbf{c}_s + \mathbf{D}_a\mathbf{c}_a \,.$$

- $\mathbf{D}_s\mathbf{c}_s = \mathbf{x}$ is the "clean" signal and $\mathbf{D}_a\mathbf{c}_a = \boldsymbol{\delta}$ is the attack signal.
- $\mathbf{D}_s$ and $\mathbf{D}_a$ are called signal and attack dictionaries.
- Index *i* corresponds to signal class and index *j* corresponds to attack class.

Formulation – Block Sparsity (contd.) and Classification

- They solve this problem in relaxed form via active set homotopy algorithm

- Signal and

$$\min_{\mathbf{c}_s,\mathbf{c}_a} \|\mathbf{x}' - \mathbf{D}_s\mathbf{c}_s - \mathbf{D}_a\mathbf{c}_a\|_2^2 + \lambda_s \sum_{i=1}^{r} \|\mathbf{c}_s[i]\|_2 + \lambda_a \sum_{i=1}^{r}\sum_{j=1}^{a} \|\mathbf{c}_a[i][j]\|_2$$

$$\hat{i} = \arg\min_i \|\mathbf{x}' - \mathbf{D}_s[i]\hat{\mathbf{c}}_s[i] - \mathbf{D}_a\hat{\mathbf{c}}_a\|_2$$

$$\hat{j} = \arg\min_j \|\mathbf{x}' - \mathbf{D}_s\hat{\mathbf{c}}_s - \mathbf{D}_a[\hat{i}][j]\hat{\mathbf{c}}_a[\hat{i}][j]\|_2$$

- They also consi ng the reconstructed image (called SBSC+CNN)

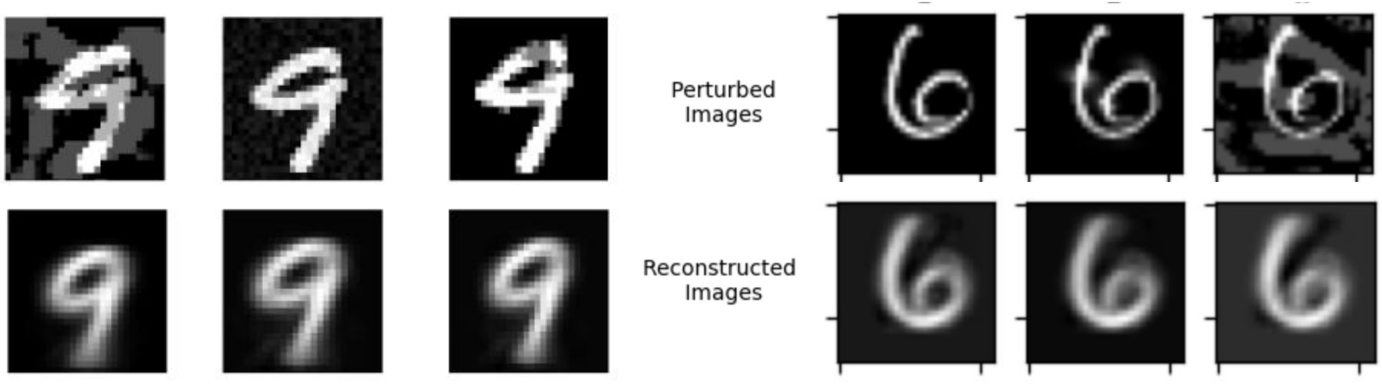
$$\hat{\mathbf{x}} = \mathbf{D}_s[\hat{i}]\mathbf{c}_s^*[\hat{i}]$$

Identification guarantees

Proposition 5.1. *The correct classes of the signal* $\mathbf{x} \in \mathcal{S}^{\mathbf{x}}_{i^*}$ *and the attack* $\boldsymbol{\delta} \in \mathcal{S}^{\boldsymbol{\delta}}_{i^*,j^*}$, *with* $\mathcal{S}^{\mathbf{x}}_{i^*} \cap \mathcal{S}^{\boldsymbol{\delta}}_{i^* j^*} = \emptyset$, *can be recovered by solving* (10) *if and only if,* $\forall i^*, j^*, \forall \mathbf{x}' \in (\mathcal{S}^{\mathbf{x}}_{i^*} \oplus \mathcal{S}^{\boldsymbol{\delta}}_{i^*,j^*}), \mathbf{x} \neq \mathbf{0}$, *the* ℓ_1/ℓ_2 *norm of the correct-class minimum* ℓ_1/ℓ_2 *vectors* $\hat{\mathbf{c}}^*_s, \hat{\mathbf{c}}^*_a$ *is strictly less that of the wrong-class minimum* ℓ_1/ℓ_2 *norm vectors* $\tilde{\mathbf{c}}^*_s, \tilde{\mathbf{c}}^*_a$, *i.e.,*

$$\|\hat{\mathbf{c}}^*_s\|_{1,2} + \|\hat{\mathbf{c}}^*_a\|_{1,2} < \|\tilde{\mathbf{c}}^*_s\|_{1,2} + \|\tilde{\mathbf{c}}^*_a\|_{1,2}. \qquad (14)$$

Sample reconstructed images from MNIST dataset



ℓ_∞ ℓ_1 ℓ_2 Attack Type ℓ_1 ℓ_2 ℓ_∞

From the paper Our reproduction

Experimental results on MNIST dataset

Table 2. Adversarial image and attack classification accuracy on digit classification of MNIST dataset. See above table for column descriptions. The clean accuracy represents the accuracy of the method with unperturbed test inputs.

MNIST	CNN	M_∞	M_2	M_1	MAX	AVG	MSD	BSC	SBSC	SBSC+CNN	SBSAD
Clean accuracy	98.99%	99.1%	99.2%	99.0%	98.6%	98.1%	98.3%	92%	94%	99%	-
ℓ_∞ PGD ($\epsilon = 0.3$)	0.03%	**90.3%**	0.4%	0.0%	51.0%	65.2%	62.7%	54%	77.27%	76.83%	73.2%
ℓ_2 PGD ($\epsilon = 2.0$)	44.13%	68.8%	69.2%	38.7%	64.1%	67.9%	70.2%	76%	**85.34%**	85.17%	46%
ℓ_1 PGD ($\epsilon = 10.0$)	41.98%	61.8%	51.1%	74.6%	61.2%	66.5%	70.4%	75%	**85.97%**	85.85%	36.6%
Average	28.71%	73.63%	40.23%	37.77%	58.66%	66.53%	67.76%	68.33%	**82.82%**	82.61%	51.93%
Unseen Attacks											
ℓ_∞ MIM ($\epsilon = 0.3$)	0.02%	**92.3%**	11.2%	0.1%	70.7%	76.7%	71.0%	59.5%	74.3%	74.2%	79.0%
ℓ_2 C-W ($\epsilon = 2.0$)	0%	79.6%	74.5%	44.8%	72.1%	72.4%	74.5%	**89.1%**	87.1%	87.1%	60.4%
ℓ_2 DDN ($\epsilon = 2.0$)	0%	63.9%	70.5%	40.0%	62.5%	64.6%	69.5%	**88.8%**	87.2%	87.1%	57.8%
Average	0%	78.6%	52.06%	28.3%	68.43%	71.23%	71.66%	79.13%	**82.86%**	82.8%	65.73%

Our Contribution

Setting

- We use the same setting as (Thaker et al., 2022)
  - Block sparse signal model
  - Active set homotopy algorithm to compute the coefficient vectors $\mathbf{c}_s$ and $\mathbf{c}_a$.
- However, (Thaker et al., 2022) does not consider detecting if an attack had occurred or not.
- Our claim is that the prediction of the original classifier pre- and post-reconstruction must be
  - Different, if the original image was attacked, and
  - Same, if it was not.

Attack Detection and Identification

- We use this claim as an attack detection scheme
- If an attack is detected, we classify it using SBSAD.

$$\hat{i} = \arg\min_{i} ||\mathbf{x}' - \mathbf{D}_s[i]\mathbf{c}_s^*[i] - \mathbf{D}_a\mathbf{c}_a^*||_2 \quad (5)$$

$$\hat{\mathbf{x}} = \mathbf{D}_s[\hat{i}]\mathbf{c}_s^*[\hat{i}] \quad (6)$$

$$\hat{j} = \arg\min_{j} ||\mathbf{x}' - \mathbf{D}_s\mathbf{c}_s^* - \mathbf{D}_a[\hat{i}][j]\mathbf{c}_a^*[\hat{i}][j]||_2 \quad (7)$$

Attack Detection and Identification Algorithm

Algorithm 1 Attack Detection and Identification Algorithm

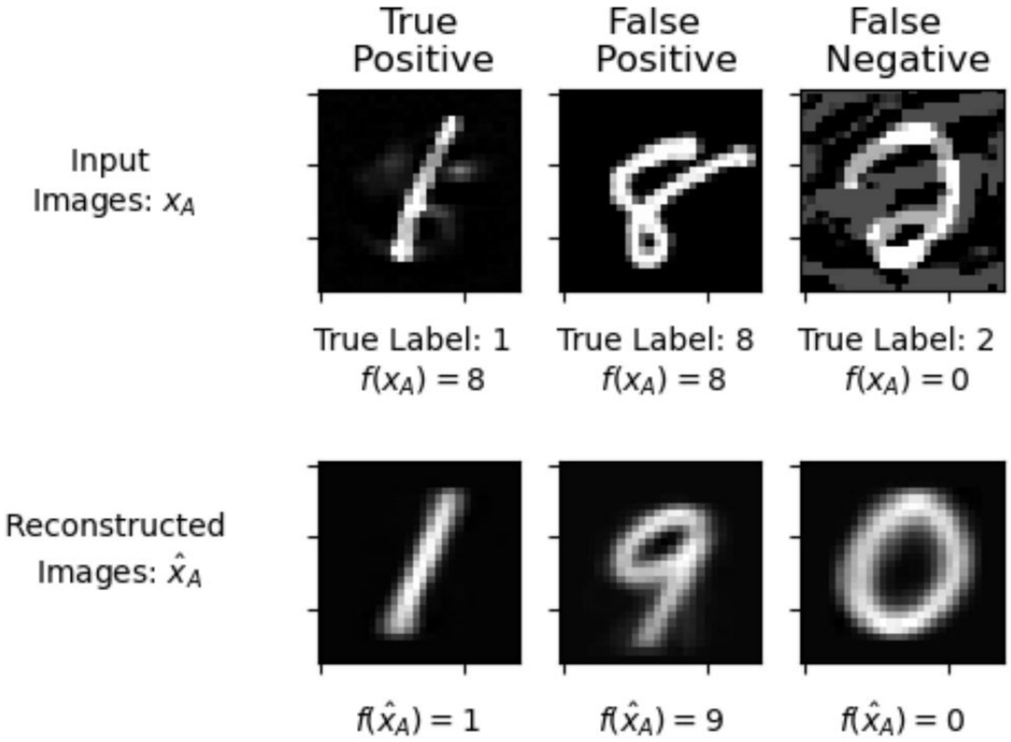
Inputs:
Corrupted Test Set: $\mathcal{X}_A$
NN Classifier: $f(\mathbf{x})$
Signal and Adversary Dictionaries: $\mathbf{D}_s,\ \mathbf{D_a}$

Initialize:
Attacked Image Set: $\mathcal{A} \leftarrow \{\}$

for all $\mathbf{x}_A \in \mathcal{X}_A$ **do**
Compute $\mathbf{c}_s^*,\ \ \mathbf{c}_a^*$ with ASHA given $\mathbf{D}_a$, $\mathbf{D}_s$, and $\mathbf{x}_A$
Compute $\hat{i}_A,\ \ \hat{\mathbf{x}}_A$, and $\hat{j}_A$ according to Eqns. (5), (6), and (7)

/* If the predictions don't match, record attack */
if $f(\mathbf{x}_A) \neq f(\hat{\mathbf{x}}_A)$ **then**
$\mathcal{A} \leftarrow \mathcal{A} \bigcup \{\left(\mathbf{x}_A,\ \hat{\mathbf{x}}_A, \hat{i}_A, \hat{j}_A\right)\}$
end if
end for
return $\mathcal{A}$

Examples of True Positives and False Positives and Negatives



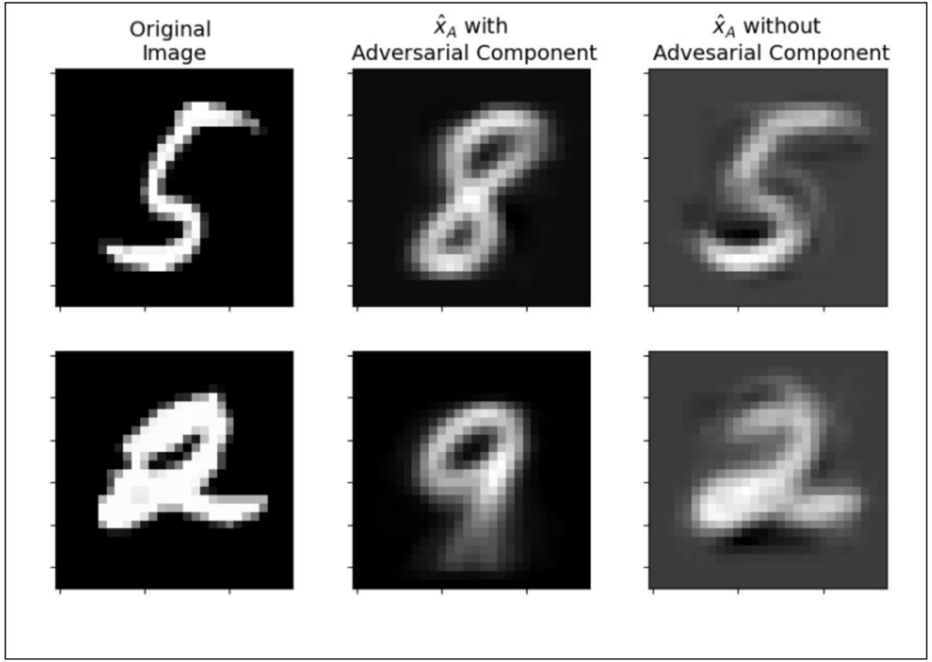
Experimental Results: Detect and Identify

Attack Parameters	Overall Detected	True Positives	False Positives	ℓ_1 **Detected**	ℓ_2 **Detected**	ℓ_∞ **Detected**	Overall Identified	ℓ_1 **Identified**	ℓ_2 **Identified**	ℓ_∞ **Identified**
$\epsilon_{\ell_1} = 5, \epsilon_{\ell_2} = 1, \epsilon_{\ell_\infty} = .15$	41 %	122	172	32 %	24 %	66 %	43 %	59 %	29 %	39 %
$\epsilon_{\ell_1} = 10, \epsilon_{\ell_2} = 2, \epsilon_{\ell_\infty} = .3$	43 %	130	117	21 %	30 %	79 %	45 %	52 %	23 %	52 %
$\epsilon_{\ell_1} = 20, \epsilon_{\ell_2} = 4, \epsilon_{\ell_\infty} = .6$	37 %	111	155	22 %	31 %	58 %	63 %	55 %	26 %	86 %

Table 1: Detection and identification results for 300 attacked images of 1000 test images from the MNIST data set with varying attack parameters. 10 images were selected at random from each class to be attack by each attack type.

Problem: How to Reduce False Positives?

Attack Parameters	Overall Detected	True Positives	False Positives
$\epsilon_{\ell_1} = 5, \epsilon_{\ell_2} = 1, \epsilon_{\ell_\infty} = .15$	41 %	122	172



Robust Attack Detection

$$\mathbf{c}_s^* = \min_{\mathbf{c}_s} \|\mathbf{x}' - \mathbf{D}_s\mathbf{c}_s\|_2^2 + \lambda_s \sum_{i=1}^{r} \|\mathbf{c}_s[i]\|_2 \,.$$

Then we reconstruct the image as follows:

$$\hat{i} = \arg\min_{i} ||\mathbf{x}' - \mathbf{D}_s[i]\mathbf{c}_s^*[i]||_2$$

$$\hat{\mathbf{x}} = \mathbf{D}_s[\hat{i}]\mathbf{c}_s^*[\hat{i}]$$

* BSC from (Thaker et al., 2022)

Robust Attack Detection Algorithm

Algorithm 2 Our Robust Attack Detection Algorithm

Inputs:

Corrupted Test Set: $\mathcal{X}_A$
NN Classifier: $f(\mathbf{x})$
Signal Dictionary: $\mathbf{D}_s$

Initialize:

Attacked Image Set: $\mathcal{A} \leftarrow \{\}$
Cleaned Image Set: $\mathcal{X}_C \leftarrow \mathcal{X}_A$

for all $\mathbf{x}_A \in \mathcal{X}_A$ **do**
    Solve (8) to get $\mathbf{c}_s^*$ with ASHA given, $\mathbf{D}_s$, and $\mathbf{x}_A$
    Compute $\hat{i}_A$, and $\hat{\mathbf{x}}_A$ according to Eqns. (9) adn (10)

    /* If the predictions don't match, record attack and remove from clean image set*/
    if $f(\mathbf{x}_A) \neq f(\hat{\mathbf{x}}_A)$ **then**
        $\mathcal{A} \leftarrow \mathcal{A} \bigcup \{\left(\mathbf{x}_A, \ \hat{\mathbf{x}}_A, \hat{i}_A\right)\}$
        $\mathcal{X}_C \leftarrow \mathcal{X}_C - \{\mathbf{x}_A\}$
    end if
end for
return $\mathcal{A}, \ \mathcal{X}_C$

Experimental Results: Robust Detection

Attack Parameters	Total Attacks	Overall Detected	True Positives (TP)	False Positives (FN)	Acc. on TP	Acc. on FN	Acc. on $\mathcal{X}_C$	Raw Acc.
$\epsilon_{\ell_1} = 5, \epsilon_{\ell_2} = 1, \epsilon_{\ell_\infty} = .15$	300	29 %	86	63	10 %	93 %	98%	89 %
$\epsilon_{\ell_1} = 10, \epsilon_{\ell_2} = 2, \epsilon_{\ell_\infty} = .3$	300	35 %	105	59	7 %	72 %	93 %	84 %
$\epsilon_{\ell_1} = 20, \epsilon_{\ell_2} = 4, \epsilon_{\ell_\infty} = .6$	300	23 %	68	48	13%	59 %	89 %	84 %

Previous result

$\epsilon_{\ell_1} = 5, \epsilon_{\ell_2} = 1, \epsilon_{\ell_\infty} = .15$	41 %	122	172
$\epsilon_{\ell_1} = 10, \epsilon_{\ell_2} = 2, \epsilon_{\ell_\infty} = .3$	43 %	130	117
$\epsilon_{\ell_1} = 20, \epsilon_{\ell_2} = 4, \epsilon_{\ell_\infty} = .6$	37 %	111	155

Experimental Results: Robust Detection

Attack Parameters	Total Attacks	Overall Detected	True Positives (TP)	False Positives (FN)	Acc. on TP	Acc. on FN	Acc. on $\mathcal{X}_C$	Raw Acc.
$\epsilon_{\ell_1} = 10, \epsilon_{\ell_2} = 2, \epsilon_{\ell_\infty} = .3$	600	34 %	205	32	5 %	77 %	88 %	72 %
$\epsilon_{\ell_1} = 10, \epsilon_{\ell_2} = 2, \epsilon_{\ell_\infty} = .3$	300	35 %	105	59	7 %	72 %	93 %	84 %
$\epsilon_{\ell_1} = 10, \epsilon_{\ell_2} = 2, \epsilon_{\ell_\infty} = .3$	150	35 %	52	58	8 %	76 %	96 %	92 %

Experimental Results: Robust Detection

Attack Parameters	Total Attacks	Overall Detected	True Positives (TP)	False Positives (FN)	Acc. on TP	Acc. on FN	Acc. on $\mathcal{X}_C$	Raw Acc.
CW	300	91 %	274	37	0 %	0 %	96 %	69%
MIM	300	60 %	180	51	0 %	20 %	87 %	72%
DDN	300	93 %	278	65	0 %	0 %	95 %	69%
Spatial Transform	300	86 %	259	66	0 %	0 %	93 %	69%

Conclusion and Future Work

- We present 2 algorithms
  - Detection and Identification
  - Robust Detection

- Many directions to explore further
  - Dict construction with core sets See Paul et al. (2021) and Mirzasoleiman et al. (2020)
  - Training on reconstructed images
  - Train on coefficient representation of images
  - Impact of targeted attacks vs untargeted
  - Scaling to larger data sets

References

Darshan Thaker, Paris Giampouras, and René Vidal. Reverse engineering ℓp attacks: A block-sparse optimization approach with recovery guarantees. In *International Conference on Machine Learning*, pp. 21253–21271. PMLR, 2022.

Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *International conference on machine learning*, pp. 274–283. PMLR, 2018.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

Baharan Mirzasoleiman, Jeff Bilmes, and Jure Leskovec. Coresets for data-efficient training of machine learning models, 2020.

Mansheej Paul, Surya Ganguli, and Gintare Karolina Dziugaite. Deep learning on a data diet: Finding important examples early in training. *Advances in Neural Information Processing Systems*, 34:20596–20607, 2021.

Thank You !