Adversarial Perturbation Attacks on GLRT-based Detectors

Ismail Alkhouri, Dr. George K. Atia, Dr. Wasfy Mikhael

DSP Lab.

Department of Electrical and Computer Engineering

University of Central Florida

Introduction

- The classification task Importance. Applications are in healthcare, finance, education, and transportation systems [Carlini et al.' 2017].
- Despite their known success, Studies have proved their vulnerability to adversarial perturbations attacks
- Attacks Categorizations based on [Akhtar et al.' 2018]:
    > side information availability: White-box, black-box, and semi black-box
    > goal of the attacker: targeted and non-targeted attacks

Introduction

- The vast majority of existing work has focused on Simple Hypothesis Testing. In this work, we formulate the problem as a Composite Hypothesis Testing

- **Contributions**:

> We derive adversarial perturbation attacks for multi-class CHT (MCHT) classifiers. The imperceptibility is gauged

> We investigate the robustness of CHT classifiers to attacks with different amount of side information acquired through additional learning mechanisms.

Adversarial Attacks: Background

For the classification of M classes, $k(.)$ be the classifier function that maps the input vector $y \in \mathbb{R}^N$ to its estimated label.

Function k is not differentiable, there exist M discriminant functionals $f_i: \mathbb{R}^N \rightarrow \mathbb{R}$, $i \in [M]$

$$k(\mathbf{y}) = \underset{i \in [M]}{\operatorname{argmax}} f_i(\mathbf{y}).$$

Let perturbations $\eta \in \mathbb{R}^N$, the classifier is fooled if: $k(\mathbf{y}) \neq k(\mathbf{y} + \eta)$.

By definition: $G(\mathbf{y} + \eta) := \min_{i \neq k(\mathbf{y})} \{ f_{k(\mathbf{y})}(\mathbf{y} + \eta) - f_i(\mathbf{y} + \eta) \} < 0$

Adversarial Attacks: Background

To be imperceptible, we need to get below optimization problem [Balda et al.' 2018]:

Find: η

$$\text{s.t. } G(\mathbf{y}) + \eta^T \nabla_y G(\mathbf{y}) < 0, \quad \|\eta\|_\infty \leq \epsilon$$

Using first-order Taylor series, we get:

$$\min_{\eta} \left\{ G(\mathbf{y}) + \eta^T \nabla_y G(\mathbf{y}) \right\} \quad \text{s.t.} \quad \|\eta\|_\infty \leq \epsilon \,. \tag{4}$$

A closed-form solution of (4) is given by,

$$d\eta = -\epsilon \, \text{sign}(\nabla_y G(\mathbf{y})) \,, \tag{5}$$

where sign(.) is the signum function.

System Model

Assume we have a LTI system with input sequence $s_i \in \mathbb{R}^L$ and unknown impulse response $\theta \in \mathbb{R}^q$. The output $x \in \mathbb{R}^N$, where $N = L + q - 1$, is observed in noise. The goal is to detect the sequence of interest $s_i, i \in [M]$, i.e.,

$$y = x + w = s_i * \theta + w, i \in [M],$$

where * denotes the discrete linear convolution.

y is the feature vector under the i^{th} hypothesis, M is the total number of hypotheses, and w is a zero-mean AWGN with covariance matrix $\sigma_w^2 I_N$. We can write the model as:

$$y = S_i\theta + w$$

System Model

S_i is the zero-padded Toeplitz matrix representing sequence S_i.

$$\mathbf{S}_i = \begin{bmatrix} \mathrm{s}_i(1) & 0 & \dots & 0 & 0 \\ \mathrm{s}_i(2) & \mathrm{s}_i(1) & 0 & \dots & 0 \\ \vdots & \mathrm{s}_i(2) & \ddots & \mathrm{s}_i(1) & \vdots \\ \mathrm{s}_i(L) & \vdots & \ddots & \mathrm{s}_i(2) & \mathrm{s}_i(1) \\ 0 & \mathrm{s}_i(L) & \ddots & \vdots & \mathrm{s}_i(2) \\ \vdots & \vdots & \vdots & \mathrm{s}_i(L) & \vdots \\ 0 & 0 & \dots & 0 & \mathrm{s}_i(L) \end{bmatrix}$$

$$J(\mathbf{y}, \theta; \mathcal{H}_i) = \|\mathbf{y} - \mathbf{S}_i\theta\|_2^2 . \tag{9}$$

Minimizing J w.r.t. θ, we obtain the MLE,

$$\theta_i^* = (\mathbf{S}_i^T \mathbf{S}_i)^\dagger \mathbf{S}_i^T \mathbf{y} , \tag{10}$$

where $(.)^\dagger$ denotes the pseudoinverse. The GLRT detector selects the hypothesis k that minimizes the cost,

$$k(\mathbf{y}) = \underset{i \in [M]}{\operatorname{argmin}} J(\mathbf{y}, \theta_i^*; \mathcal{H}_i). \tag{11}$$

Proposed Solution

We present two methods to generate perturbations:

1- In our composite setting, we use the negative cost function: $-\nabla_y J_{k(\mathbf{y})}(\mathbf{y})$.

2- The function $G(y+\eta)$ is replaced by $J_{i'}(y+\eta) - J_{k(y)}(y+\eta)$, and

$$\hat{i} = \underset{i \neq k(\mathbf{y})}{\operatorname{argmin}} \frac{|J_i(\mathbf{y}) - J_{k(\mathbf{y})}(\mathbf{y})|}{||\nabla_y J_i(\mathbf{y}) - \nabla_y J_{k(\mathbf{y})}(\mathbf{y})||_1} . \qquad (12)$$

Perturbation examples are then generated by solving (4). The closed-form solution is obtained as

$$\eta = -\epsilon \ \operatorname{sign}(\nabla_y J_{\hat{i}}(\mathbf{y}) - \nabla_y J_{k(\mathbf{y})}(\mathbf{y})) . \qquad (13)$$

Proposed Solution

Given that the above two methods depend on calculating the gradient of the cost function, we define the $N \times N$ matrix:

$$\mathbf{\Gamma}_i = \mathbf{S}_i(\mathbf{S}_i^T\mathbf{S}_i)^\dagger\mathbf{S}_i^T$$

Thus,

$$\nabla_y J_i(\mathbf{y}) = 2(\mathbf{I}_N - \mathbf{\Gamma}_i)(\mathbf{y} - \mathbf{\Gamma}_i\mathbf{y}) \ .$$

Experimental Results:

We present the performance of the proposed methods for generating perturbations in two cases. (Case 1) is when the attacker has no prior knowledge about the ground truth label, therefore has to obtain the estimated label. (Case 2) is when attacker has prior knowledge of the ground truth label through some learning mechanism . Both cases fall under the category of white-box non-targeted attacks.

$$\zeta = \frac{CA - CA_{pert}}{CA}$$

$$\rho_p = \frac{||\eta||_p}{||\mathbf{y}||_p}$$

Experimental Setup:

In our experiments:

- s_i is selected from discrete uniform distribution over $\{-0.5, 0.5\}$
- θ and w are selected from Gaussian distributions of zero mean and I_q and $0.25 I_N$ variance.
- The length of sequences L = 10,
- The number of unknown parameters q = 30, and the number of hypotheses M = 15.
- Results are averaged over 1000 trials

Experimental Results:

TABLE I: Fooling ratio ζ^{method} as a function of ϵ and ρ_p.

ϵ	0.05	0.07	0.09	0.11	0.13	0.15	0.17	0.19	0.21	0.23	0.25	0.27	0.29	0.31
$\rho_2(\%)$	3.51	4.95	6.34	7.72	9.23	10.71	12.12	13.45	14.77	16.13	17.64	18.93	20.38	21.78
$\rho_\infty(\%)$	1.43	2.02	2.57	3.12	3.78	4.34	4.92	5.48	6.06	6.54	7.17	7.67	8.31	8.89
$\zeta^1(\%)$ of Case 1	4.93	11.68	13.19	20.86	26.11	34.92	37.16	41.70	49.23	55.39	60.17	63.84	64.98	71.26
$\zeta^1(\%)$ of Case 2	13.19	19.92	25.03	33.45	39.60	45.17	50.49	56.53	64.34	67.78	71.81	75.85	76.32	81.09
$\zeta^2(\%)$ of Case 1	11.09	21.40	27.00	35.55	42.57	56.77	57.90	61.68	67.00	72.73	75.00	75.36	78.54	80.59
$\zeta^2(\%)$ of Case 2	22.56	33.82	41.43	52.83	59.52	69.96	74.93	80.15	87.05	90.08	90.80	92.27	93.95	96.26

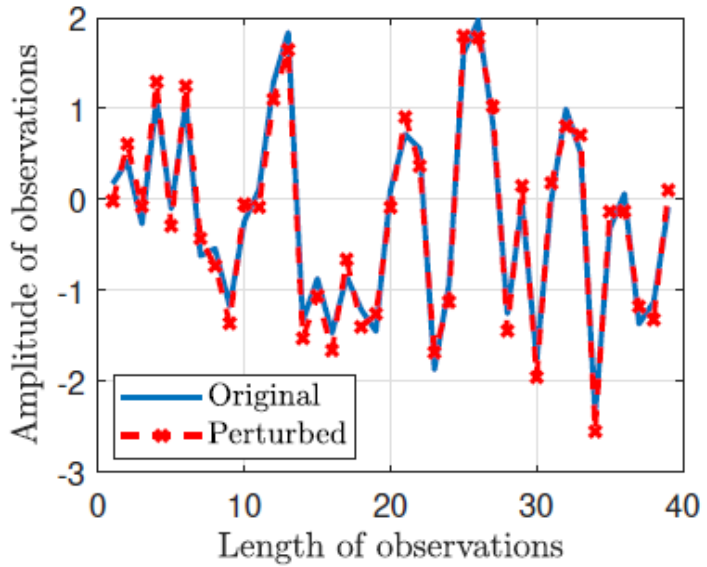

Fig. 1: Original observation and perturbed version of Method 1 and Case 2.

Experimental Results:

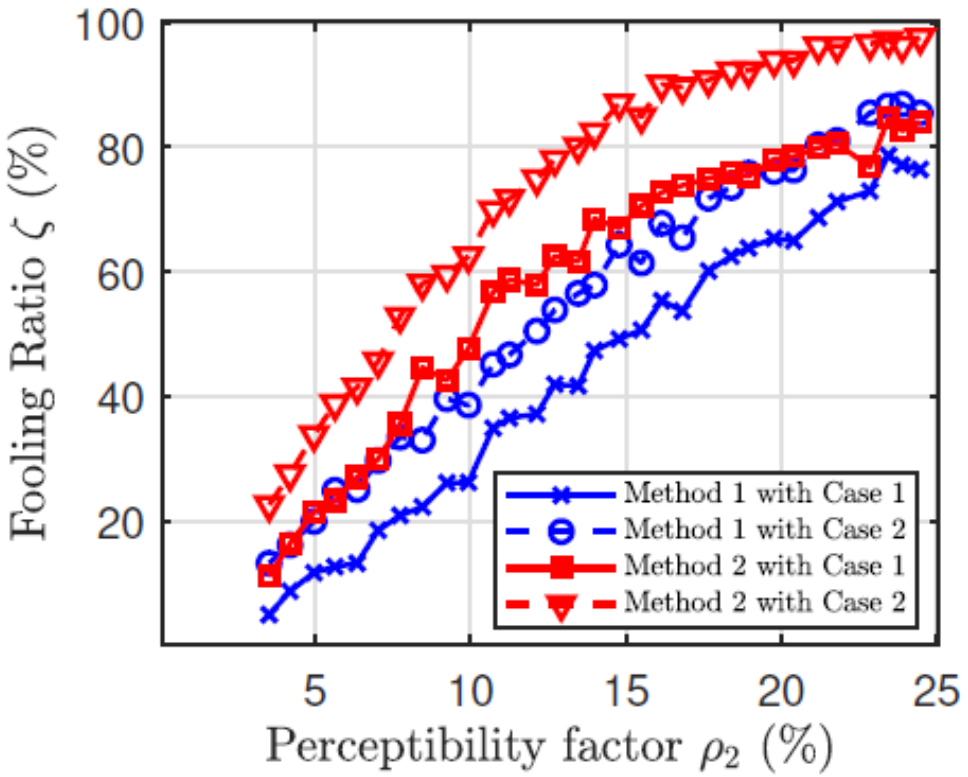

Fig. 3: Fooling ratio vs. Perceptibility factor

TABLE II: Run-time results with $\epsilon = 0.19$, $CA = 81\%$, and $\rho_2 = 13.45\%$.

Method	Case	$CA_{pert}(\%)$	$\zeta(\%)$	run-time (seconds)
1	1	47.28	45.70	13.10
1	2	35.25	56.53	12.87
2	1	31.08	61.68	19.57
2	2	16.10	80.15	19.57

TABLE III: Results with various input symbols.

Input	CA	CA_{pert}	$\zeta(\%)$	ϵ	$\rho_2(\%)$	$\rho_\infty(\%)$
4-PAM	65.70	5.5	91.63	0.19	16.45	6.91
$\mathcal{U}(-0.5, 0.5)$	45.80	2.1	95.78	0.15	16.47	6.6
$\mathcal{N}(0, \mathbf{I})$	96.9	26.7	72.45	0.42	16.56	6.74

References

[Carlini et al.’ 2017] Carlini, Nicholas, and David Wagner. "Towards evaluating the robustness of neural networks." 2017 ieee symposium on security and privacy (sp). IEEE, 2017.

[Akhtar et al.’ 2018] Akhtar, Naveed, and Ajmal Mian. "Threat of adversarial attacks on deep learning in computer vision: A survey." IEEE Access 6 (2018): 14410-14430.

[Balda et al.’ 2018] Balda, Emilio Rafael, Arash Behboodi, and Rudolf Mathar. "On generation of adversarial examples using convex programming." 2018 52nd Asilomar Conference on Signals, Systems, and Computers. IEEE, 2018.

ACKNOWLEDGEMENT

This work was supported in part by NSF CAREER Award

CCF-1552497.

Thank You