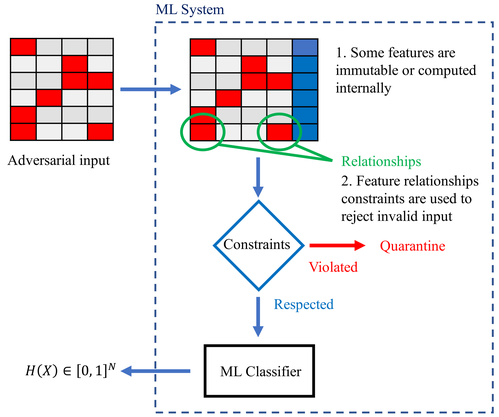
Constrained Adaptive Attack: Effective Adversarial Attack Against Deep Neural Networks for Tabular Data

Thibault Simonetto[1], Salah Ghamizi[2,3] and Maxime Cordy[1]
[1]University of Luxembourg, [2]Luxembourg institute of Science and Technologies, [3]Riken AIP
Contact: thibault.simonetto@uni.lu

Problem



Attacker objective

For x find δ such that:

- ✓ $H(x) \neq H(x+\delta)$
- ✓ $L_p(x, x+\delta) < \epsilon$
- ✓ $x + \delta \in \mathcal{X}_\Omega$

With:

$H: \mathcal{X} \to \mathcal{Y}$ the classification model

L_p the distance according to a $p-norm$

$\mathcal{X}_\Omega \subseteq \mathcal{X}$ the subspace where x satisfies all constraints $\omega \in \Omega$

Example of constraint:

$$installment = loan_amount \times \frac{int_rate \times (1 + int_rate)^{term}}{(1 + int_rate)^{term} - 1}$$

Constraints grammar:

$$\omega := \omega_1 \wedge \omega_2 \mid \omega_1 \vee \omega_2 \mid \psi_1 \succeq \psi_2$$

$$\psi := c \mid f_i \mid \psi_1 \oplus \psi_2 \mid x_i$$

Known attacks produce invalid examples:

PGD ✗

AutoAttack ✗

	Max trans.	Avg trans.	Acc. creation	Age
Adversarial	\$2500	\$3000	1 year	22

Adversarial violates Avg $transaction \leq Max\ transaction$

Constraints as a penalty function:

Constraint	Penalty function	Constraint example	Penalty function example
$\psi_1 = \psi_2$	$\mid \psi_1 - \psi_2 \mid$	rec_per_month = record/month	$\mid$ rec_per_month − (record/month) $\mid$
$\psi_1 \leq \psi_2$	$max(0, \psi_1 - \psi_2)$	open_acc ≤ total_acc	$max(0,$ open_acc − total_acc$)$
$\psi_1 < \psi_2$	$max(0, \psi_1 - \psi_2 + \tau)$	open_acc < total_acc	$max(0,$ open_acc − total_acc $+10^{-5})$
$\omega_1 \wedge \omega_2$	$\omega_1 + \omega_2$	((term = 36) ∨ (term = 60))∧ (open_acc ≤ total_acc)	min($\mid$ term − 36 $\mid$, $\mid$ term − 60 $\mid$)+ $max(0,$ open_acc − total_acc$)$
$\omega_1 \vee \omega_2$	$\min(\omega_1, \omega_2)$	(term = 36) ∨ (term = 60)	min($\mid$ term − 36 $\mid$, $\mid$ term − 60 $\mid$)

Method

CAPGD (white-box)

Projected Gradient Descent (PGD)

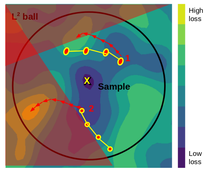
$$x^{(k+1)} = P_{\mathcal{S}}(x^{(k)} + \eta \nabla l(h(x), y))$$

Gradient loss + Constraints regularization

$$\mathcal{L}'(x) = l(h(x), y) - \sum_{\omega_i \in \Omega} penalty(x, \omega_i)$$

Constrained Gradient Descent

$$z^{(k+1)} = P_{\mathcal{S}}(x^{(k)} + \eta^{(k)}(\nabla \mathcal{L}'(x^{(k)}))$$



Adaptive Step size η

$$\eta^{(0)} = 2\epsilon,\ \eta^{(k+1)} = \begin{cases} \eta^{(k)}/2, & \text{if } \mathcal{L}' \text{ does not decrease} \\ \eta^{(k)}, & \text{otherwise} \end{cases}$$

Gradient step momentum α

$$x^{(k+1)} = R_\Omega(P_{\mathcal{S}}(x^{(k)} + \alpha \cdot (z^{(k+1)} - x^{(k)}) \quad \text{Current gradient direction}$$
$$+(1-\alpha) \cdot (x^{(k)} - x^{(k+1)})) \quad \text{Previous perturbation direction}$$

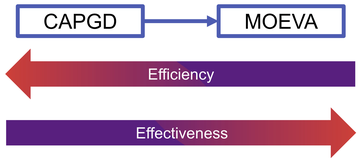
Repair operator R_Ω

$$\omega_1 \equiv installment = loan_amount \times \frac{int_rate \times (1 + int_rate)^{term}}{(1 + int_rate)^{term} - 1}$$

Threat model: Requires model's parameters and gradients

Constrained Adaptive Attack (CAA)

Sequential attack:



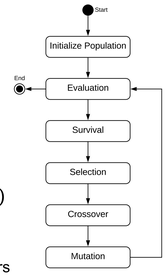
MOEVA (gray-box)

Multi-objective genetic algorithm (NSGA-III)

$$minimise\ g_1(x) \equiv h(x)$$

$$minimise\ g_2(x) \equiv L_p(x - x_0)$$

$$minimise\ g_3(x) \equiv \sum_{\omega_i \in \Omega} penalty(x, \omega_i)$$



Threat model: Does not require model's parameters or gradients

Metrics

Adversarial *examples* = ***Perturbed*** *examples if satisfied constraints*
Revert to ***Original*** *examples otherwise*

$$\textbf{\textit{Robust}}\ accuracy = \frac{\#\ Correctly\ classified\ \textbf{\textit{adversarial}}\ examples}{\#\ Clean\ examples}$$

Cost = Execution time

Experimental Settings

Three domains, four datasets

Credit scoring Lending Club Loan Data (LCLD) | Botnet detection CTU | URL phishing URL | ICU survival WIDS

Five neural network architectures

- 2 Regularizations based
- 2 Transformers based
- 1 Self and Semi-supervised based

Two training methods

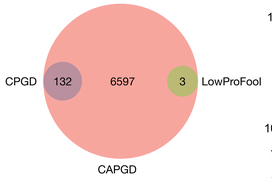
- Standard
- Madry adversarial training

Six tabular attacks

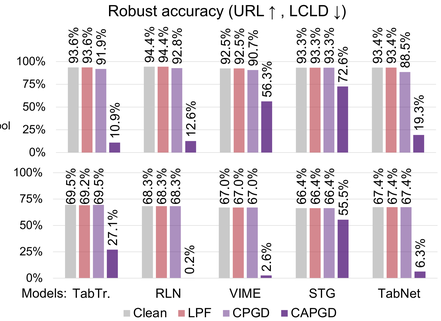
- LowProFool (gradient) ✗
- CPGD (gradient) ✗
- CAPGD (gradient) ✓
- BF* (search) ✓
- MOEVA (search) ✓
- CAA (mixed) ✓

Results

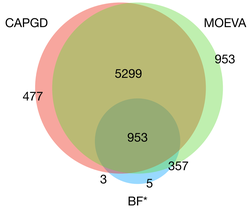
CAPGD



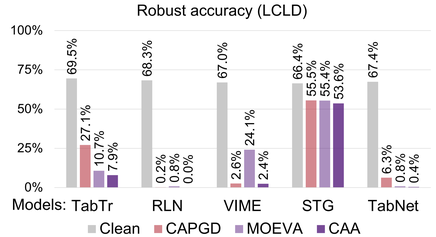
Set and number of successfull attacks across all models and datasets



CAA



Set and number of successfull attacks across all models and datasets



Robust accuracy of clean models

Dataset	Model	Clean	CAPGD	BF*	MOEVA	CAA
		Robust accuracy (↓)				
URL	TabTr.	93.6	$10.9_{\pm 0.1}$	$93.2_{\pm 0}$	$18.2_{\pm 0.8}$	$\mathbf{8.9_{\pm 0.2}}$
	RLN	94.4	$12.6_{\pm 0.2}$	$93.8_{\pm 0}$	$23.6_{\pm 0.5}$	$\mathbf{10.8_{\pm 0.2}}$
	VIME	92.5	$56.3_{\pm 0.1}$	$92.2_{\pm 0}$	$56.5_{\pm 0.9}$	$\mathbf{49.5_{\pm 0.5}}$
	STG	93.3	$72.6_{\pm 0.0}$	$93.2_{\pm 0}$	$58.2_{\pm 0.9}$	$\mathbf{58.0_{\pm 0.8}}$
	TabNet	93.4	$19.3_{\pm 0.6}$	$90.9_{\pm 0}$	$17.5_{\pm 0.6}$	$\mathbf{11.0_{\pm 0.5}}$
LCLD	TabTr.	69.5	$27.1_{\pm 0.9}$	$61.1_{\pm 0}$	$10.7_{\pm 0.8}$	$\mathbf{7.9_{\pm 0.6}}$
	RLN	68.3	$0.2_{\pm 0.1}$	$38.9_{\pm 0}$	$0.8_{\pm 0.2}$	$\mathbf{0.0_{\pm 0.0}}$
	VIME	67.0	$2.6_{\pm 0.2}$	$52.6_{\pm 0}$	$24.1_{\pm 1.5}$	$\mathbf{2.4_{\pm 0.1}}$
	STG	66.4	$55.5_{\pm 0.2}$	$\mathbf{53.0_{\pm 0}}$	$55.4_{\pm 0.2}$	$53.6_{\pm 0.1}$
	TabNet	67.4	$6.3_{\pm 0.4}$	$49.0_{\pm 0}$	$0.8_{\pm 0.1}$	$\mathbf{0.4_{\pm 0.1}}$
CTU	TabTr.	95.3	$\mathbf{95.3_{\pm 0.0}}$	$95.3_{\pm 0}$	$95.3_{\pm 0.0}$	$95.3_{\pm 0.0}$
	RLN	97.8	$97.8_{\pm 0.0}$	$97.5_{\pm 0}$	$\mathbf{94.0_{\pm 0.2}}$	$\mathbf{94.0_{\pm 0.2}}$
	VIME	95.1	$95.1_{\pm 0.0}$	$95.1_{\pm 0}$	$\mathbf{40.8_{\pm 4.7}}$	$\mathbf{40.8_{\pm 4.7}}$
	STG	95.3	$95.3_{\pm 0.0}$	$95.3_{\pm 0}$	$95.3_{\pm 0.0}$	$95.3_{\pm 0.0}$
	TabNet	96.1	$96.1_{\pm 0.0}$	$13.0_{\pm 0}$	$\mathbf{0.0_{\pm 0.0}}$	$\mathbf{0.0_{\pm 0.0}}$
WIDS	TabTr.	75.5	$48.0_{\pm 0.3}$	$67.7_{\pm 0}$	$59.2_{\pm 0.6}$	$\mathbf{45.9_{\pm 0.3}}$
	RLN	77.5	$61.8_{\pm 0.3}$	$77.0_{\pm 0}$	$67.7_{\pm 0.3}$	$\mathbf{60.9_{\pm 0.2}}$
	VIME	72.3	$51.4_{\pm 0.3}$	$71.2_{\pm 0}$	$59.4_{\pm 0.5}$	$\mathbf{50.3_{\pm 0.2}}$
	STG	77.6	$65.1_{\pm 0.4}$	$77.5_{\pm 0}$	$68.8_{\pm 0.3}$	$\mathbf{63.8_{\pm 0.2}}$
	TabNet	79.7	$10.2_{\pm 0.3}$	$73.1_{\pm 0}$	$13.9_{\pm 0.4}$	$\mathbf{5.3_{\pm 0.4}}$

Robust accuracy of adversarially trained models

Dataset	Accuracy	TabTr.	RLN	VIME	STG	TabNet
		Robust accuracy (↓)				
URL	Clean	$93.9_{\pm 0.3}$	$95.2_{\pm 0.8}$	$93.4_{\pm 0.9}$	$94.3_{\pm 1.0}$	$99.5_{\pm 6.1}$
	CAA	$56.7_{\pm 47.8}$	$56.2_{\pm 45.4}$	$69.8_{\pm 20.3}$	$90.0_{\pm 32.0}$	$91.8_{\pm 80.8}$
LCLD	Clean	$73.9_{\pm 4.4}$	$69.5_{\pm 1.2}$	$65.5_{-1.5}$	$15.6_{-50.8}$	$0.0_{-67.4}$
	CAA	$70.3_{\pm 62.4}$	$63.0_{\pm 63.0}$	$10.4_{\pm 8.0}$	$12.1_{-41.5}$	$0.0_{-0.4}$
CTU	Clean	$95.3_{\pm 0.0}$	$97.3_{-0.5}$	$95.1_{\pm 0.0}$	$95.1_{-0.2}$	$0.2_{-95.8}$
	CAA	$95.3_{\pm 0.0}$	$97.1_{\pm 3.0}$	$94.0_{\pm 53.2}$	$95.1_{-0.2}$	$0.2_{\pm 0.2}$
WIDS	Clean	$77.3_{\pm 1.8}$	$78.0_{\pm 0.5}$	$72.1_{-0.2}$	$62.6_{-15.1}$	$98.4_{\pm 18.6}$
	CAA	$65.1_{\pm 19.2}$	$66.6_{\pm 5.7}$	$52.1_{\pm 1.8}$	$45.2_{-18.6}$	$58.4_{\pm 53.3}$

Learn more

Paper | Code

This project is supported by the Luxembourg National Research Fund and BGL BNP Paribas Luxembourg.