

NeurIPS 2024

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

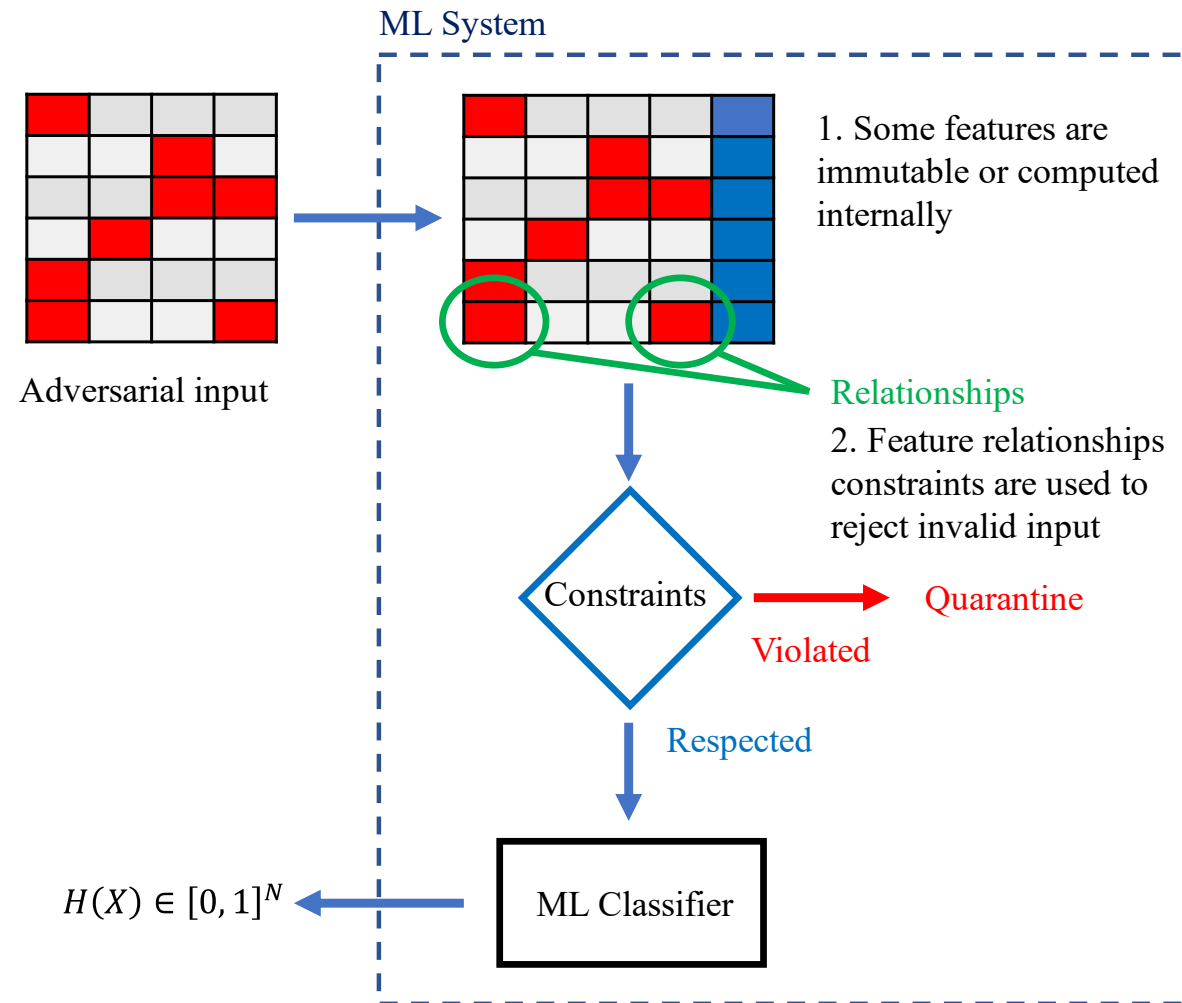
Thibault Simonetto¹, Salah Ghamizi^{2, 3}, Maxime Cordy¹

¹University of Luxembourg, Luxembourg

²Luxembourg Institute of Science and Technology, Luxembourg

³RIKEN Center for Advanced Intelligence Project, Tokyo, Japan

Adversarial examples in tabular data



Relation constraints on feature space

Finance:

avg transaction amount $\leq$ max transaction amount

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

Problem formulation

Given a classification model H ,
 a maximum perturbation ϵ under a L_p distance
 a set of **constraints** Ω

Objective of **constrained** adversarial attacks, for clean sample x find perturbation δ :

✓ With $H(x) \neq H(x + \delta)$

✓ With $L_p(x, x + \delta) < \epsilon$

✓ $x + \delta \models \Omega$

Feature relation constraints as a penalty function

Objective

$penalty(x, \omega)$ how far is x from satisfying ω $penalty(x, \omega) = 0$ iff $\omega(x) = \mathbf{True}$

Example

$\omega_1 \equiv avg_transaction \leq max_transaction$ $penalty(x, \omega_1) = max(0, avg_transaction - max_transaction)$

Mapping to penalty functions

Constraint	Penalty function
$\psi_1 = \psi_2$	$ \psi_1 - \psi_2 $
$\psi_1 \leq \psi_2$	$max(0, \psi_1 - \psi_2)$
$\psi_1 < \psi_2$	$max(0, \psi_1 - \psi_2 + \tau)$
$\omega_1 \wedge \omega_2$	$\omega_1 + \omega_2$
$\omega_1 \vee \omega_2$	$\min(\omega_1, \omega_2)$

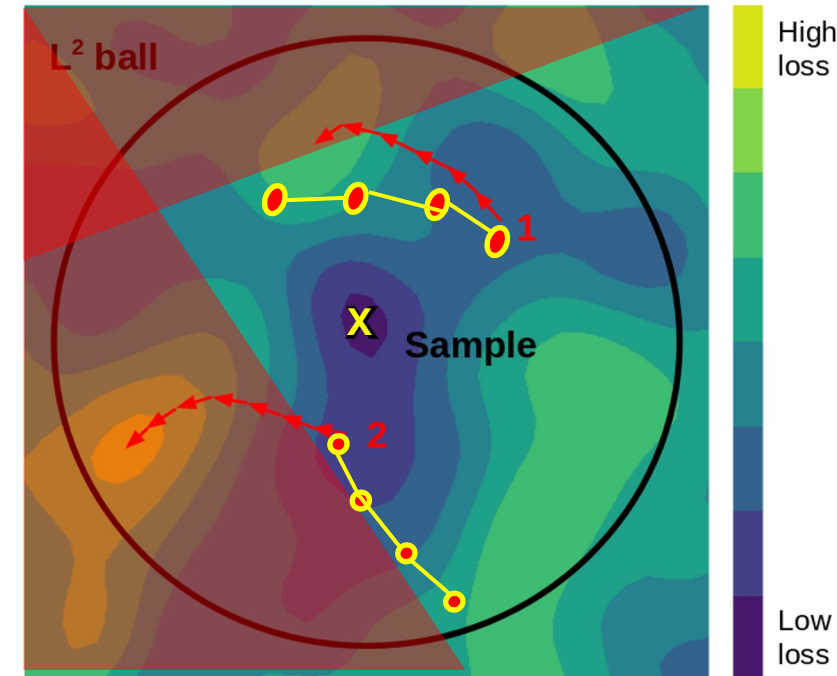
Constrained Adaptive PGD

Projected Gradient Descent (PGD)

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

Gradient loss + Constraints regularization

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



Constrained Adaptive PGD

Gradient loss + Constraints regularization

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

Constrained Gradient Descent

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

Adaptive Step size η

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

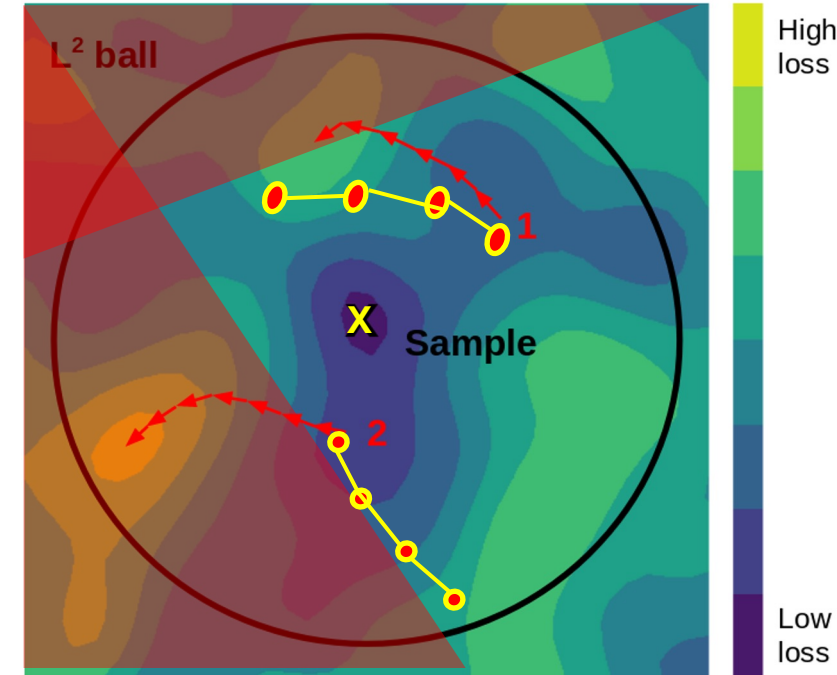
Gradient step momentum α

$$x^{(k+1)} = R_{\Omega}(P_{\mathcal{S}}(x^{(k)} + \alpha \cdot (z^{(k+1)} - x^{(k)}) + (1 - \alpha) \cdot (x^{(k)} - x^{(k-1)})))$$

Current gradient direction
Previous perturbation direction

Repair operator R_{Ω}

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



Experimental settings

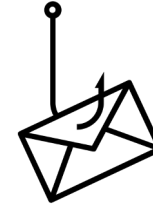
Datasets:



Credit scoring
Lending Club Loan Data



Botnet detection
CTU



URL phishing
URL

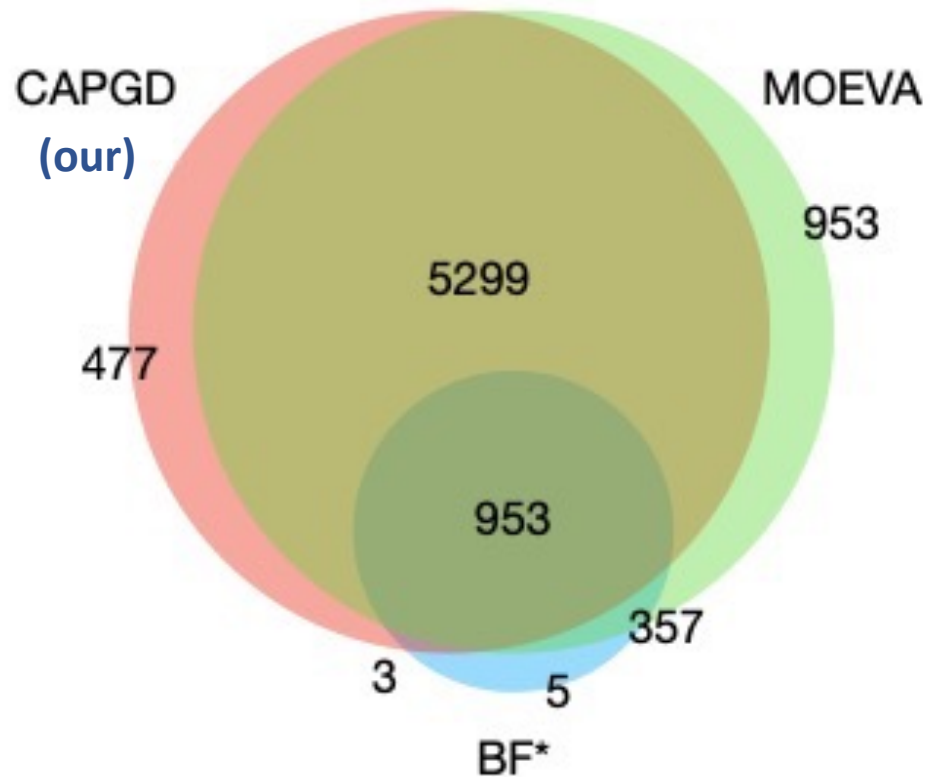


ICU survival
WIDS

Models: 5 Neural network architectures

- 2 Regularizations
- 2 Transformers
- 1 Semi-supervised

Are these attacks complementary?



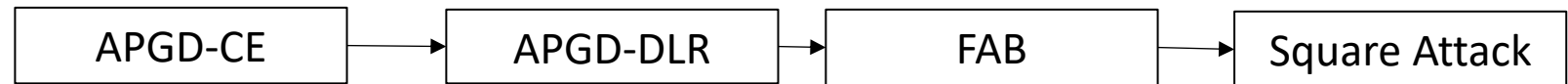
Set and number of successful attacks
across all models and datasets

Insight:

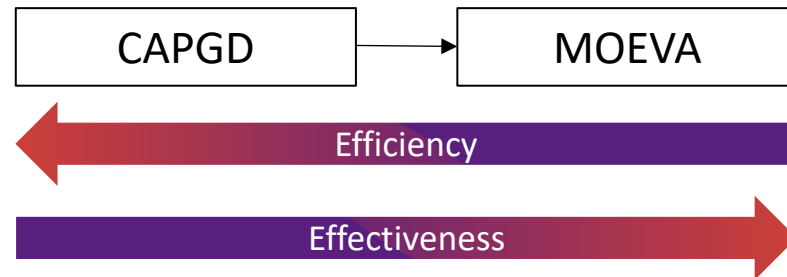
1. MOEVA and CAPGD are complementary
2. Together they subsume BF*

Constrained Adaptive Attack

Inspired from AutoAttack [1]



We propose CAA

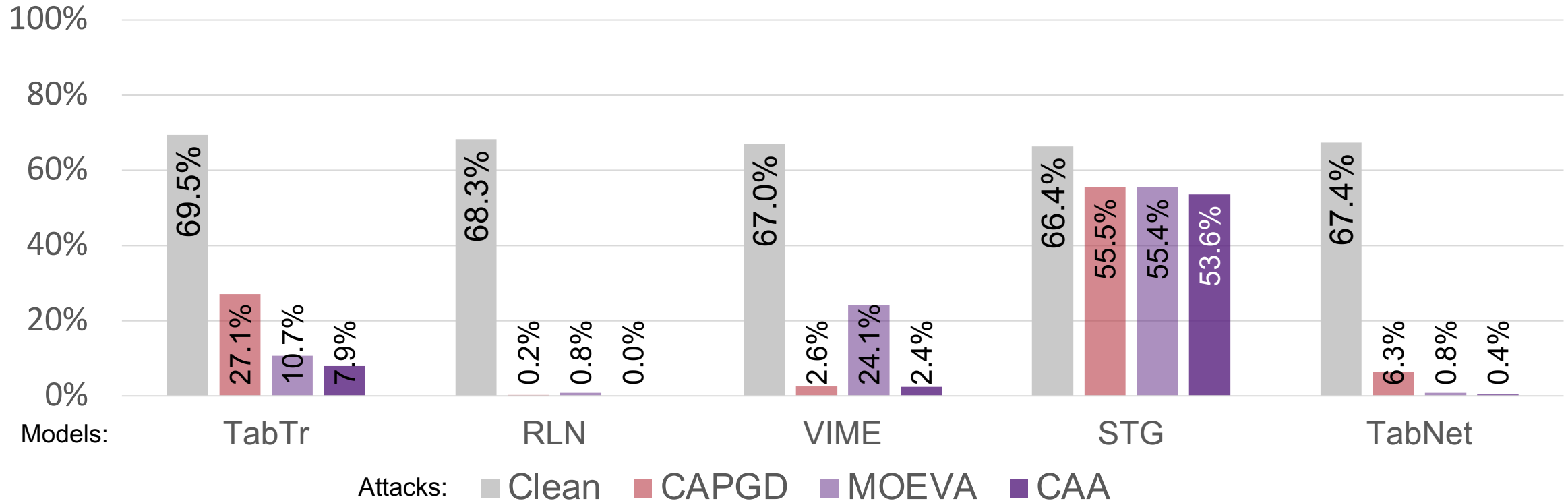


99.9%

Of examples can be generated by CAA

Impact of CAA

Robust accuracy



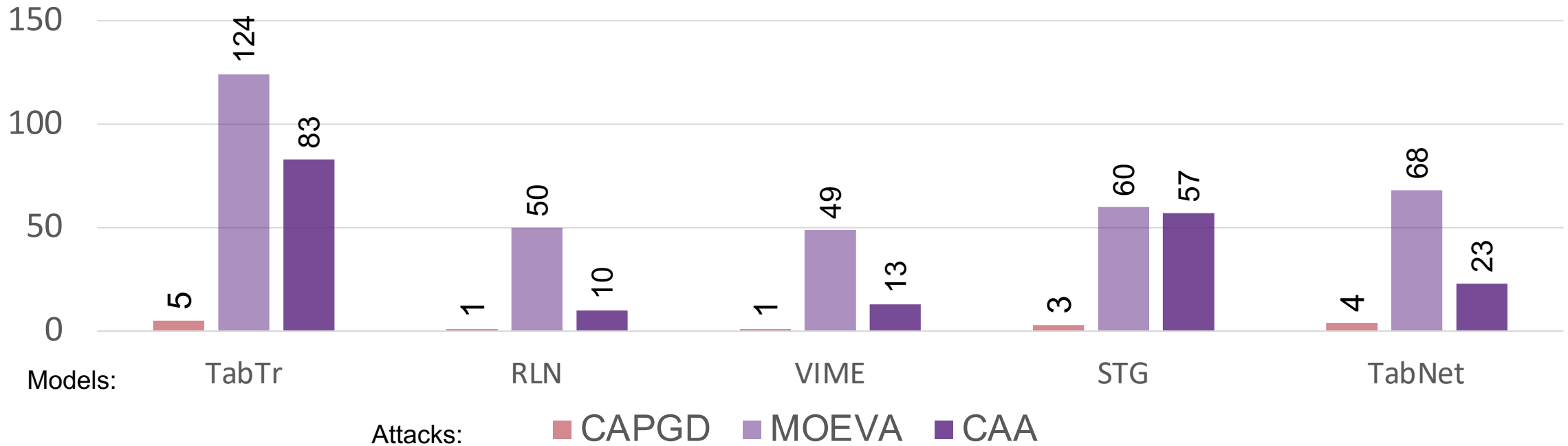
Dataset: Lending Club Loan data

Insight:

CAA effectively combines the benefits of CAPGD and MOEVA

Efficiency of CAA

Duration in seconds



Dataset: Lending Club Loan data

Insight:
CAA reduces execution costs by up to 5 times compared to MOEVA

Conclusion

RobustBench

ROBUSTBENCH Leaderboards Paper

Available Leaderboards

CIFAR-10 (ℓ_∞) CIFAR-10 (ℓ_2) CIFAR-10 (Corruptions) CIFAR-100 (ℓ_∞) ImageNet (ℓ_∞) ImageNet (Corruptions: IN-C, IN-3DCC)

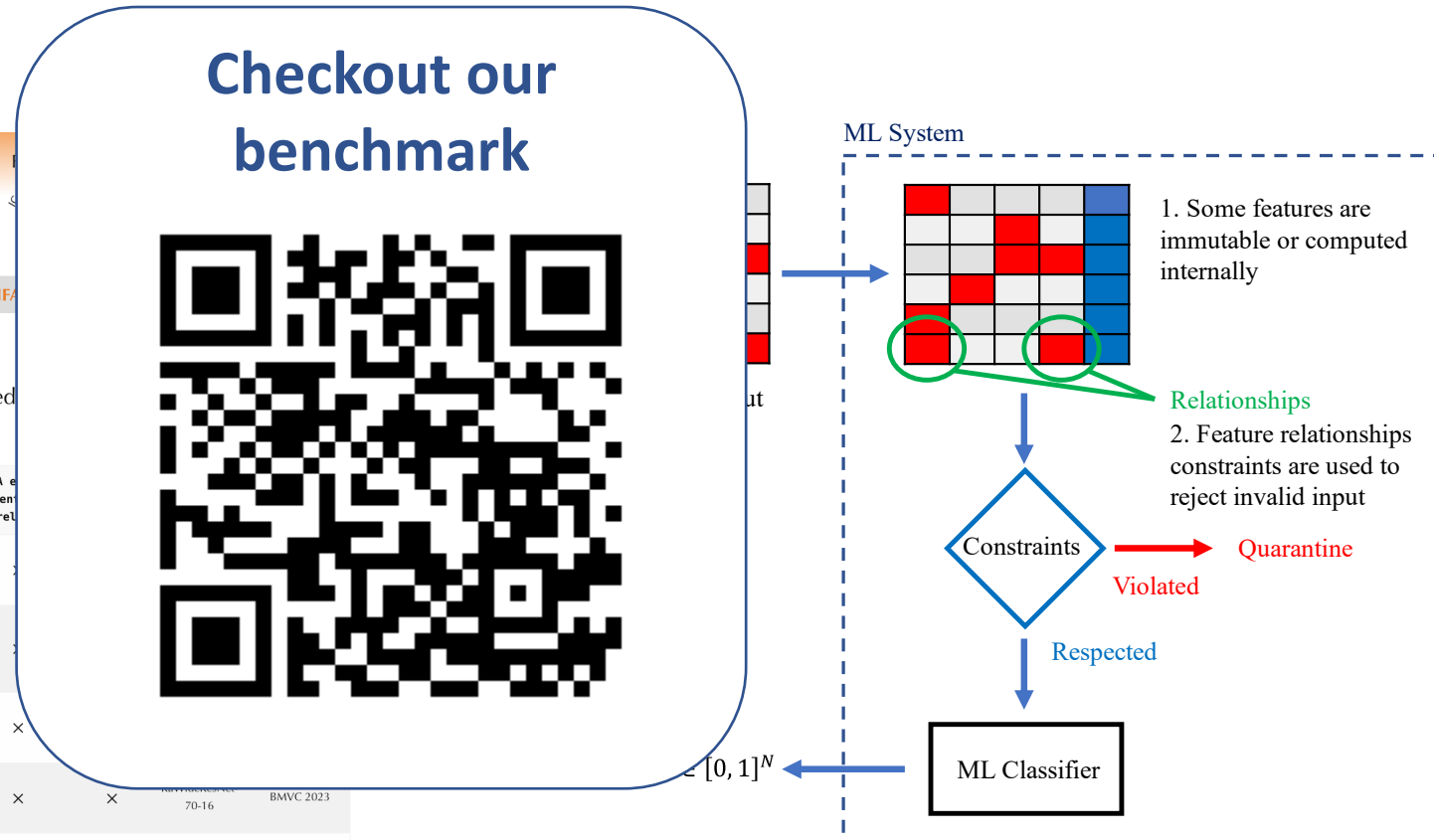
Leaderboard: CIFAR-10, $\ell_\infty = 8/255$, untargeted

Show 15 entries

Rank	Method	Standard accuracy	AutoAttack robust accuracy	Best known robust accuracy	AA e potent unrel
1	Adversarial Robustness Limits via Scaling-Law and Human-Alignment Studies <i>It uses additional 300M synthetic images in training.</i>	93.68%	73.71%	73.71%	
2	MeanSparse: Post-Training Robustness Enhancement Through Mean-Centered Feature Sparsification <i>It adds the MeanSparse operator to the adversarially trained models.</i>	93.24%	72.08%	72.08%	
3	Adversarial Robustness Limits via Scaling-Law and Human-Alignment Studies <i>It uses additional 300M synthetic images in training.</i>	93.11%	71.59%	71.59%	×
4	Robust Principles: Architectural Design Principles for Adversarially Robust CNNs <i>It uses additional 50M synthetic images in training.</i>	93.27%	71.07%	71.07%	×
5	Better Diffusion Models Further Improve Adversarial Training <i>It uses additional 50M synthetic images in training.</i>	93.25%	70.69%	70.69%	×

<https://robustbench.github.io/>

CAA: tabular attack under constraints



<https://github.com/serval-uni-lu/tabularbench>