MAX-PLANCK-INSTITUT
FÜR BIOLOGISCHE KYBERNETIK

Pluribus One
seeing one in many

Lab
Pattern Recognition
and Applications Lab

University of
Cagliari, Italy

Fast minimum-norm adversarial attacks through adaptive norm constraints

Maura Pintor, Fabio Roli, Wieland Brendel, Battista Biggio

Thirty-fifth Conference on Neural Information Processing Systems

NEURAL INFORMATION
PROCESSING SYSTEMS

Introduction

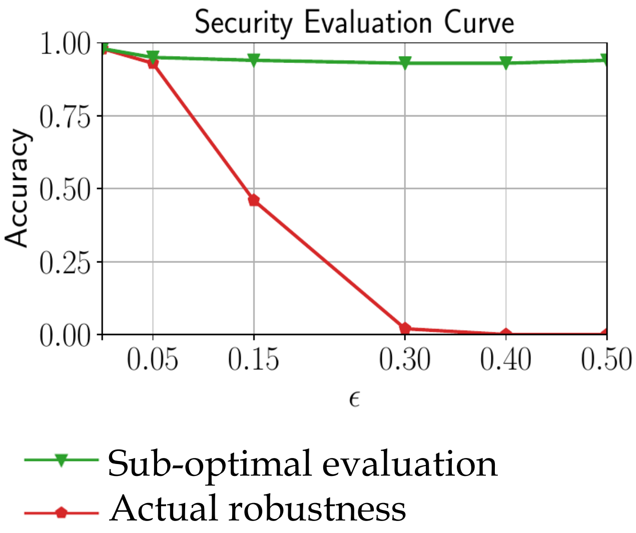
class: 9

adversarial
perturbation

class: 4

$$\boldsymbol{\delta}^\star \in \arg\min_{\boldsymbol{\delta}} \quad \|\boldsymbol{\delta}\|_p,$$
$$s.t. \quad L(\boldsymbol{x}+\boldsymbol{\delta}, y, \boldsymbol{\theta}) < 0,$$
$$\boldsymbol{x}+\boldsymbol{\delta} \in [0,1]^d,$$

Challenges of evaluating adversarial robustness



The optimization depends on the **points** and on the **model** under attack

The risk of using sub-optimal hyperparameters might lead to **over-optimistic evaluations** (failing attacks)

Available algorithms do not usually maintain stable performances for different points

Most advanced attacks aim to obtain better solutions, at the cost of **longer execution**

Adversarial perturbations with minimum norm

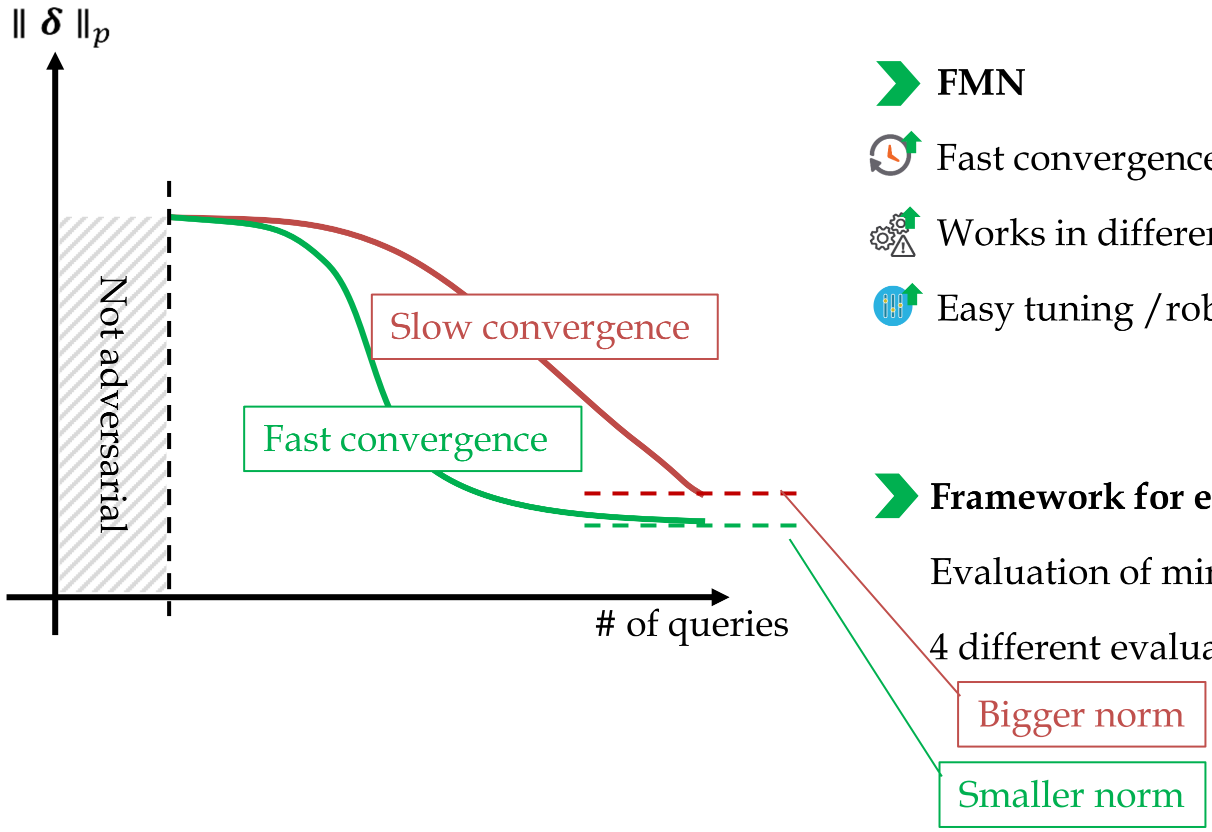
- **Carlini-Wagner attack (CW)**
  - Requires many steps to converge
- **Brendel&Bethge attack (BB)**
  - Needs initialization
  - Suffers from poor initialization
  - Complicated steps
- **Fast Adaptive Boundary (FAB)**
  - Complicated steps
  - Only untargeted version
- **Decoupling Direction & Norm (DDN)**
  - Specific to L2 norm

Long runtime

Sensitive to hyperparameters

Limited threat model

Our contributions



FMN

- Fast convergence to good local optima
- Works in different norms ($\ell_0, \ell_1, \ell_2, \ell_\infty$)
- Easy tuning / robust to hyperparameter choice

Framework for evaluating existing attacks

Evaluation of minimum-norm attacks

4 different evaluation criteria

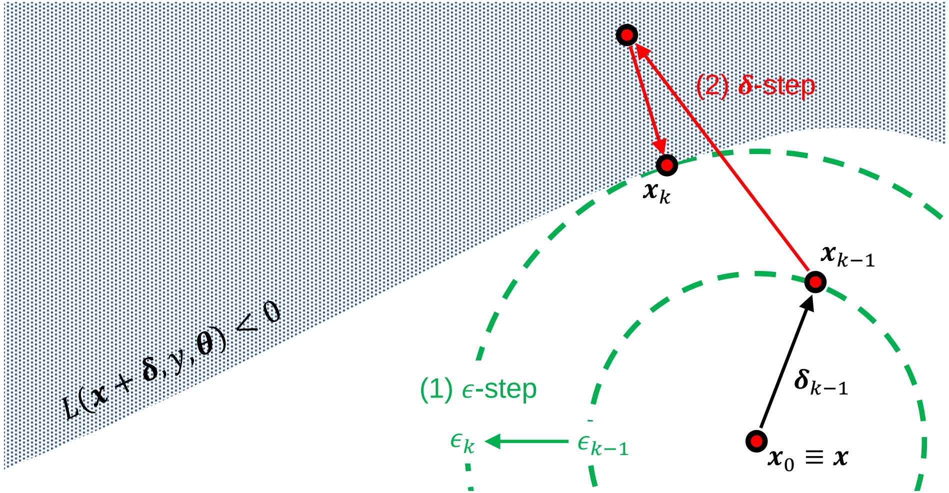
Our attack: Fast Minimum-Norm (FMN)

Algorithm 1 Fast Minimum-norm (FMN) Attack

Input: $\boldsymbol{x}$, the input sample; t, a variable denoting whether the attack is targeted ($t = +1$) or untargeted ($t = -1$); y, the target (true) class label if the attack is targeted (untargeted); γ_0 and γ_K, the initial and final ϵ-step sizes; α_0 and α_K, the initial and final $\boldsymbol{\delta}$-step sizes; K, the total number of iterations.

Output: The minimum-norm adversarial example $\boldsymbol{x}^\star$.

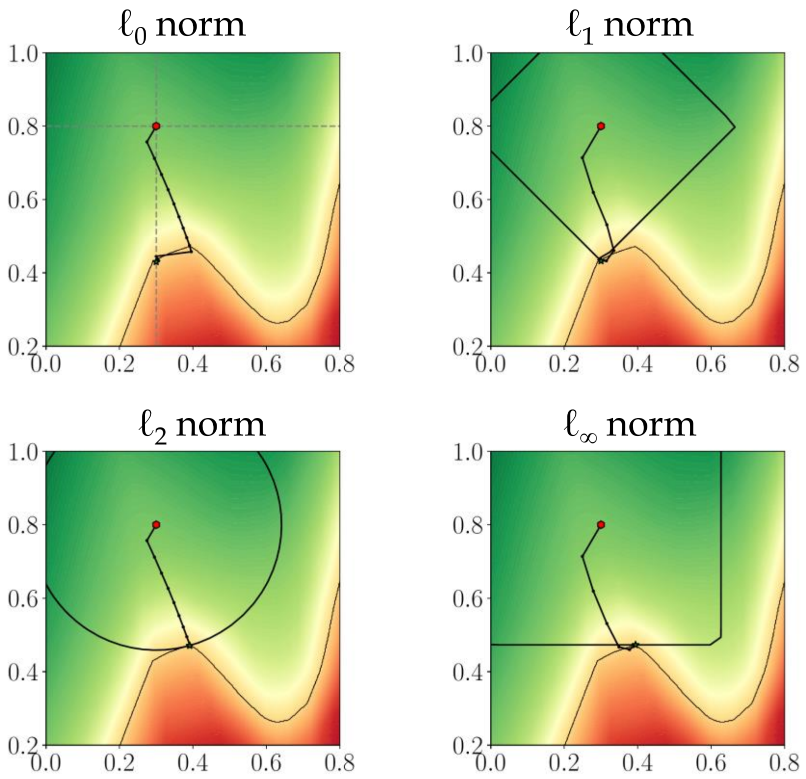
$\boldsymbol{x}_0 \leftarrow \boldsymbol{x}$, $\epsilon_0 = 0$, $\boldsymbol{\delta}_0 \leftarrow \mathbf{0}$, $\boldsymbol{\delta}^\star \leftarrow \infty$
for $k = 1, \dots, K$ **do**
  $\boldsymbol{g} \leftarrow t \cdot \nabla_{\boldsymbol{\delta}} L(\boldsymbol{x}_{k-1} + \boldsymbol{\delta}, y, \boldsymbol{\theta})$ *// loss gradient*
  $\gamma_k \leftarrow h(\gamma_0, \gamma_K, k, K)$ *// ϵ-step size decay* (Eq. 7)
  if $L(\boldsymbol{x}_{k-1}, y, \boldsymbol{\theta}) \geq 0$ **then**
    $\epsilon_k = \|\boldsymbol{\delta}_{k-1}\|_p + L(\boldsymbol{x}_{k-1}, y, \boldsymbol{\theta})/\|\boldsymbol{g}\|_q$ **if** *adversarial not found yet* **else** $\epsilon_k = \epsilon_{k-1}(1 + \gamma_k)$
  else
    if $\|\boldsymbol{\delta}_{k-1}\|_p \leq \|\boldsymbol{\delta}^\star\|_p$ **then**
      $\boldsymbol{\delta}^\star \leftarrow \boldsymbol{\delta}_{k-1}$ *// update best min-norm solution*
    end if
    $\epsilon_k = \min(\epsilon_{k-1}(1 - \gamma_k), \|\boldsymbol{\delta}^\star\|_p)$
  end if
  $\alpha_k \leftarrow h(\alpha_0, \alpha_K, k, K)$ *// $\boldsymbol{\delta}$-step size decay* (Eq. 7)
  $\boldsymbol{\delta}_k \leftarrow \boldsymbol{\delta}_{k-1} + \alpha_k \cdot \boldsymbol{g}/\|\boldsymbol{g}\|_2$
  $\boldsymbol{\delta}_k \leftarrow \Pi_\epsilon(\boldsymbol{x}_0 + \boldsymbol{\delta}_k) - \boldsymbol{x}_0$
  $\boldsymbol{\delta}_k \leftarrow \text{clip}(\boldsymbol{x}_0 + \boldsymbol{\delta}_k) - \boldsymbol{x}_0$
  $\boldsymbol{x}_k \leftarrow \boldsymbol{x}_0 + \boldsymbol{\delta}_k$
end for
return $\boldsymbol{x}^\star \leftarrow \boldsymbol{x}_0 + \boldsymbol{\delta}^\star$



Perturbation models

FMN can find minimum-norm perturbations in 4 different ℓ_p norms

In each iteration, the attack performs a step in the direction of the gradient, and then projects the point back into the ℓ_p-ball of the corresponding norm

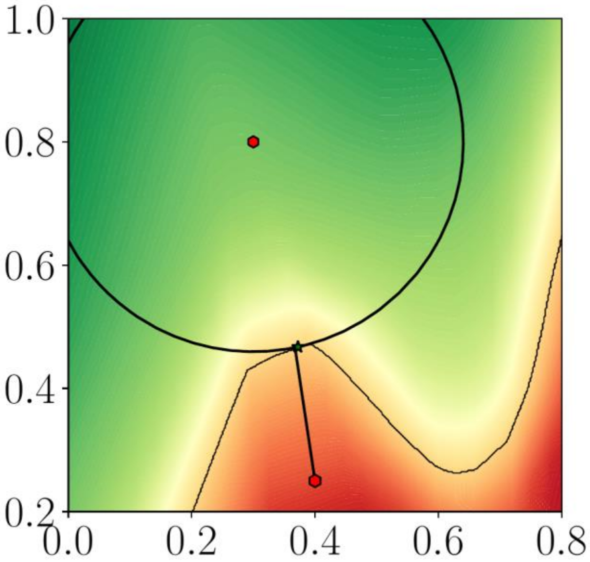


Adversarial Initialization

FMN can be initialized from a point in the target class

Finds the boundary quickly with an initial line search

Refines the results with the remaining iterations



Experimental setup

- 1 standard training model + 7 robust models in MNIST and CIFAR10 datasets (+2 ImageNet models)
- Comparison against 4 state-of-the-art minimum-distance attacks
- Targeted and untargeted scenario
- Evaluation across 4 ℓ_p distances ($p \in \{0, 1, 2, \infty\}$)

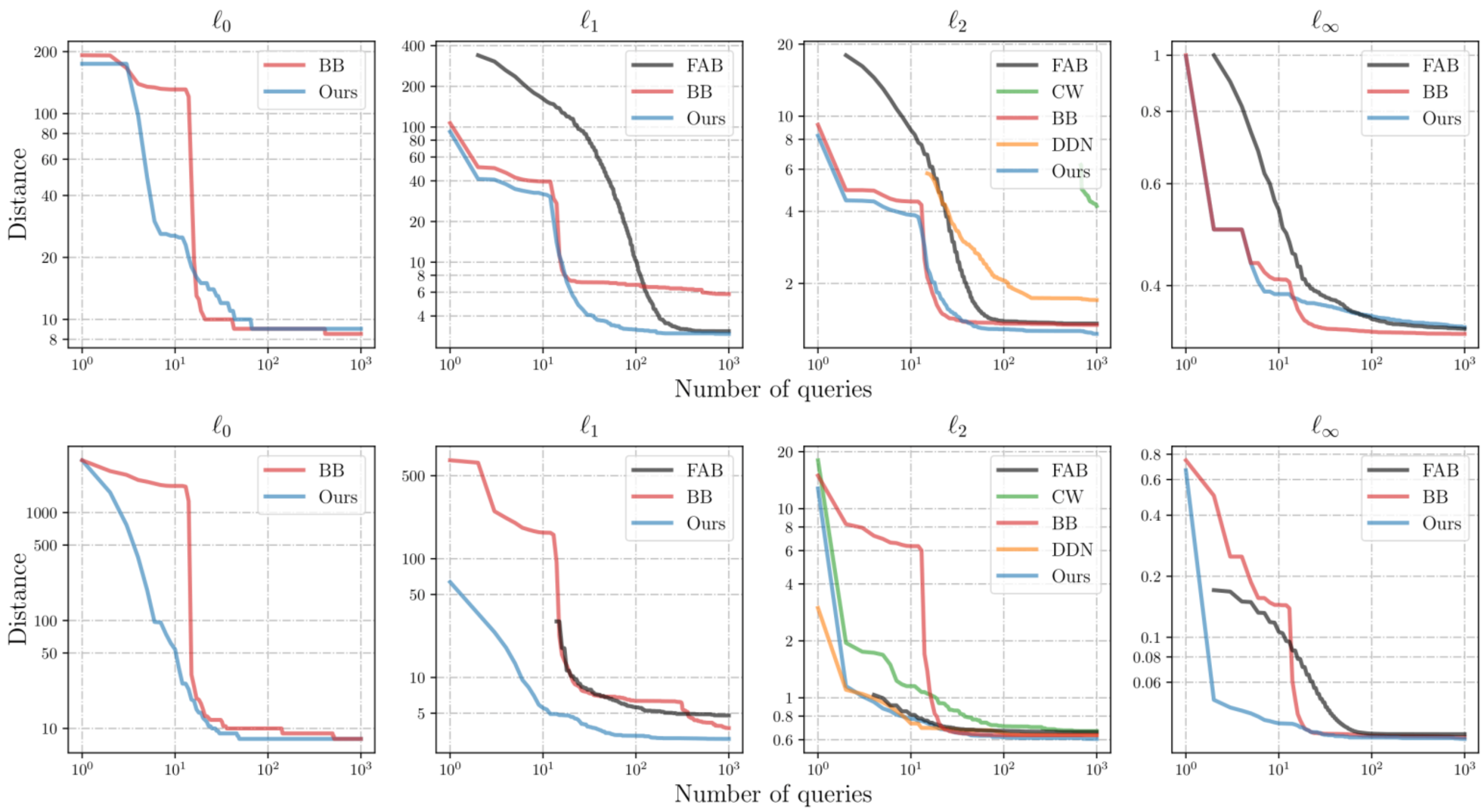
Evaluation

- Norm of perturbation
- Time-efficiency
- Query-efficiency
- Robustness to hyperparameters choice

Query-distortion curves

MNIST challenge

CIFAR challenge



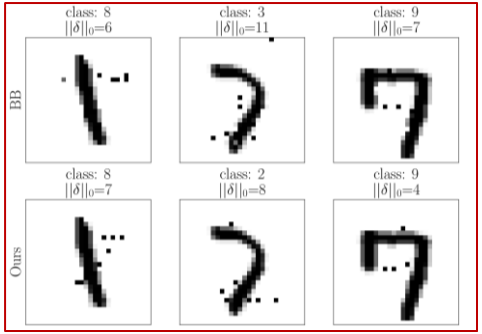
Time efficiency

		Mnist Targeted	Mnist Untargeted	Cifar Targeted	Cifar Untargeted
ℓ_0	BB	61.79 ± 0.79	10.93 ± 0.84	102.84 ± 2.85	49.08 ± 2.23
	Ours	$\mathbf{5.24} \pm \mathbf{0.56}$	$\mathbf{5.30} \pm \mathbf{0.51}$	$\mathbf{29.07} \pm \mathbf{2.56}$	$\mathbf{29.23} \pm \mathbf{2.58}$
ℓ_1	FAB	—	10.29 ± 2.02	—	100.53 ± 14.28
	BB	43.49 ± 0.23	7.02 ± 0.29	72.12 ± 2.73	35.85 ± 2.85
	Ours	$\mathbf{5.54} \pm \mathbf{0.50}$	$\mathbf{5.56} \pm \mathbf{0.49}$	$\mathbf{29.34} \pm \mathbf{2.89}$	$\mathbf{29.90} \pm \mathbf{2.22}$
ℓ_2	FAB	—	11.27 ± 1.89	—	100.86 ± 14.37
	CW	4.50 ± 0.56	4.50 ± 0.59	29.46 ± 3.19	29.51 ± 3.13
	BB	26.73 ± 0.52	4.54 ± 0.45	52.39 ± 3.16	30.12 ± 3.01
	DDN	$\mathbf{3.69} \pm \mathbf{0.54}$	$\mathbf{3.69} \pm \mathbf{0.54}$	$\mathbf{27.58} \pm \mathbf{3.44}$	$\mathbf{27.75} \pm \mathbf{3.13}$
	Ours	4.80 ± 0.58	4.79 ± 0.61	28.55 ± 2.74	28.39 ± 3.04
ℓ_∞	FAB	—	11.83 ± 1.92	—	101.34 ± 14.48
	BB	37.96 ± 1.36	14.71 ± 1.48	86.75 ± 2.71	62.11 ± 0.68
	Ours	$\mathbf{4.62} \pm \mathbf{0.60}$	$\mathbf{4.62} \pm \mathbf{0.58}$	$\mathbf{28.33} \pm \mathbf{3.08}$	$\mathbf{28.33} \pm \mathbf{3.03}$

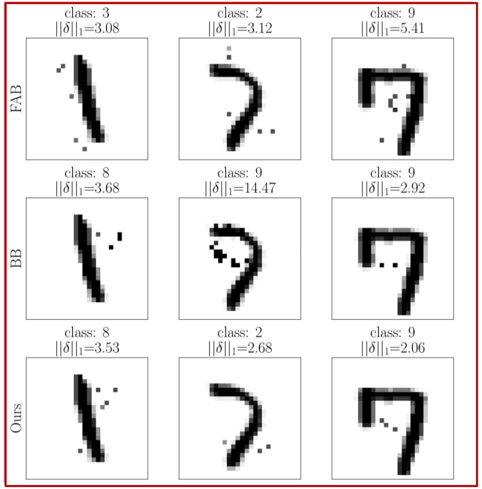
- Faster queries
- Easy tuning of the hyperparameters (more results in the paper)

Adversarial examples

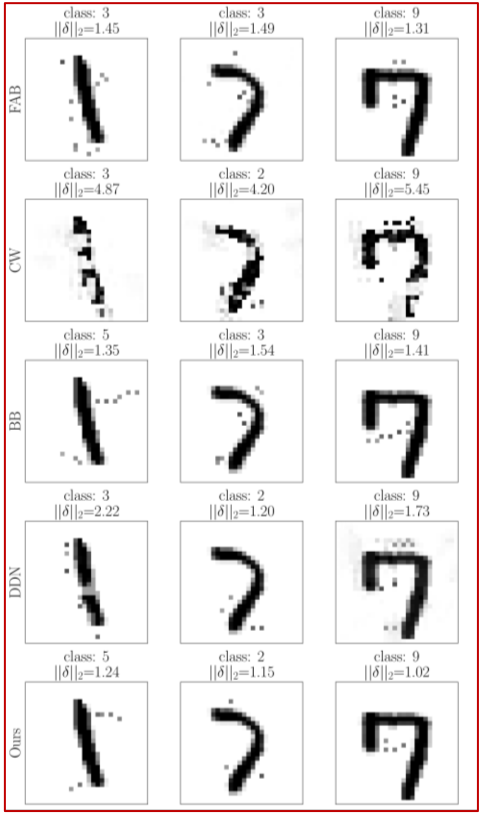
ℓ_0



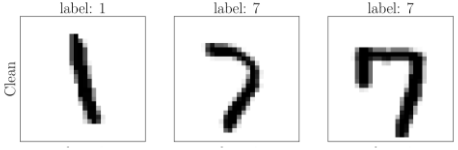
ℓ_1



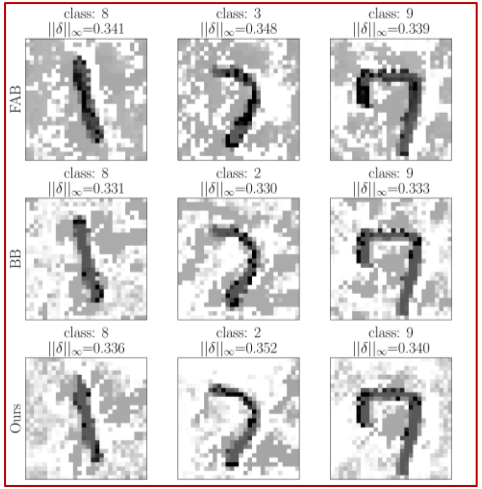
ℓ_2

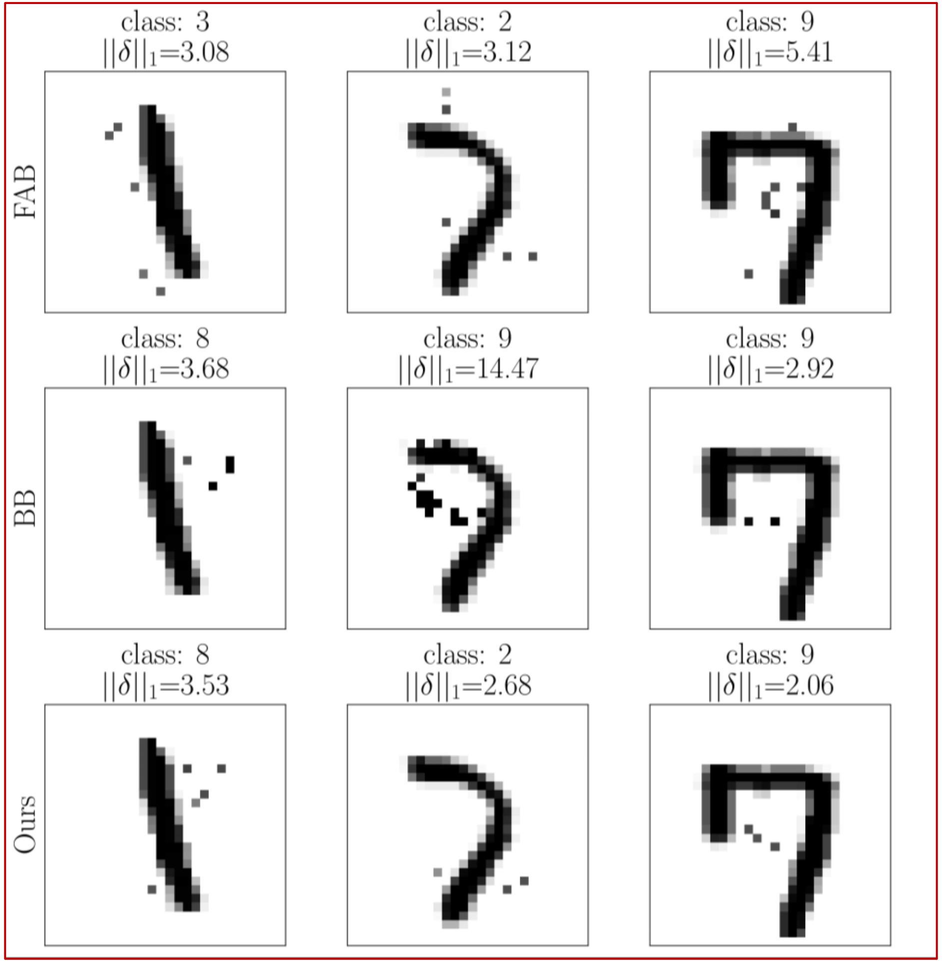


Clean



ℓ_∞





class: 3
$||\delta||_1$=3.08
class: 2
$||\delta||_1$=3.12
class: 9
$||\delta||_1$=5.41
FAB
class: 8
$||\delta||_1$=3.68
class: 9
$||\delta||_1$=14.47
class: 9
$||\delta||_1$=2.92
BB
class: 8
$||\delta||_1$=3.53
class: 2
$||\delta||_1$=2.68
class: 9
$||\delta||_1$=2.06
Ours

Conclusions

Fast-Minimum Norm Attack

- Works in different norms
- Comparable or better norm of perturbation
- Query- and time-efficient
- Robust to hyperparameter choices

We provide

- Extensive experiments
- Open-source code

Available implementations:

- https://github.com/pralab/Fast-Minimum-Norm-FMN-Attack
- https://github.com/bethgelab/foolbox
- https://github.com/jeromerony/adversarial-library
- https://github.com/pralab/secml

MAX-PLANCK-INSTITUT FÜR BIOLOGISCHE KYBERNETIK

Pluribus One
seeing one in many

Pattern Recognition and Applications Lab

University of Cagliari, Italy

Fast minimum-norm adversarial attacks through adaptive norm constraints

Thank you for listening!