Fast minimum-norm adversarial attacks through adaptive norm constraints

Maura Pintor, Fabio Roli, Wieland Brendel, Battista Biggio

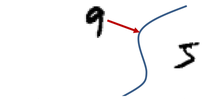
Thirty-fifth Conference on Neural Information Processing Systems (NeurIPS 2021)

Pattern Recognition and Application Laboratory
Università degli Studi di Cagliari
Pluribus One — seeing one in many
Max-Planck-Institut für Biologische Kybernetik

Neural Information Processing Systems

Adversarial Robustness Evaluation

Evaluating adversarial robustness amounts to finding the minimum norm perturbation to have input samples misclassified

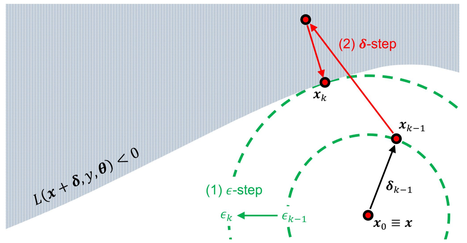


$$\delta^* \in \arg\min_{\delta} \; \| \delta \|_p, \quad \text{s.t.} \quad L(x+\delta, y, \theta) < 0, \; x+\delta \in [0,1]^d,$$

Limitations of state-of-the-art attacks

- they require many iterations to converge to good local optima;
- they require careful and computationally-demanding hyperparameter tuning; and
- they are specific to a given perturbation model.

Fast Minimum-norm Adversarial Attack

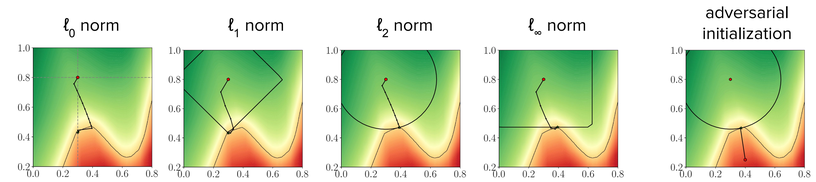


ϵ-step (1): decision-based adaptation of the search region (increase radius if x_i is not adversarial, reduce if x_i is adversarial)

δ-step (2): follows gradient direction and projects in the search region

Key aspects

Works in different ℓ_p norms



Attack Algorithm

Algorithm 1 Fast Minimum-norm (FMN) Attack

Input: 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 δ-step sizes; K, the total number of iterations.

Output: The minimum-norm adversarial example x^*.

$x_0 \leftarrow x$, $\epsilon_0 = 0$, $\delta_0 \leftarrow \mathbf{0}$, $\delta^* \leftarrow \infty$
for $k = 1, \ldots, K$ **do**
  $g \leftarrow t \cdot \nabla_\delta L(x_{k-1} + \delta, y, \theta)$ *// loss gradient*
  $\gamma_k \leftarrow h(\gamma_0, \gamma_K, k, K)$ *// ϵ-step size decay* (Eq. 7)
  if $L(x_{k-1}, y, \theta) \geq 0$ **then**
    $\epsilon_k = \|\delta_{k-1}\|_p + L(x_{k-1}, y, \theta)/\|g\|_q$ **if** *adversarial not found yet* **else** $\epsilon_k = \epsilon_{k-1}(1+\gamma_k)$
  else
    if $\|\delta_{k-1}\|_p \leq \|\delta^*\|_p$ **then**
      $\delta^* \leftarrow \delta_{k-1}$ *// update best min-norm solution*
    end if
    $\epsilon_k = \min(\epsilon_{k-1}(1-\gamma_k), \|\delta^*\|_p)$
  end if
  $\alpha_k \leftarrow h(\alpha_0, \alpha_K, k, K)$ *// δ-step size decay* (Eq. 7)
  $\delta_k \leftarrow \delta_{k-1} + \alpha_k \cdot g/\|g\|_2$
  $\delta_k \leftarrow \Pi_\epsilon(x_0 + \delta_k) - x_0$
  $\delta_k \leftarrow \text{clip}(x_0 + \delta_k) - x_0$
  $x_k \leftarrow x_0 + \delta_k$
end for
return $x^* \leftarrow x_0 + \delta^*$

Experiments

Experimental setup

- tested on 9 models (MNIST, CIFAR10, ImageNet)
- compared against 4 state-of-the-art minimum-norm attacks
- evaluated in targeted and untargeted scenario

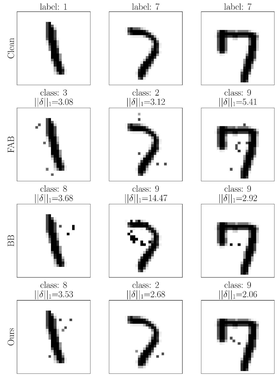
Evaluation metrics

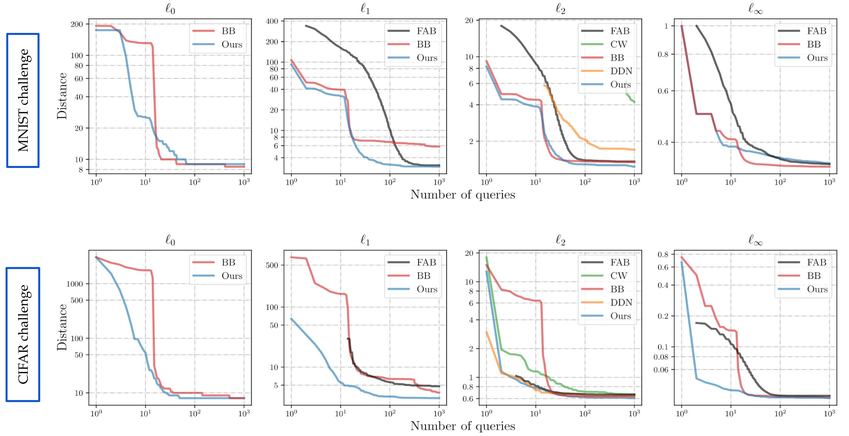
- median distance found after 1000 queries
- avg. time per query
- convergence speed
- robustness to hyperparameter choice

Median norm of perturbation $\|\delta^*\|_p$ found for 1000 queries

		MNIST								CIFAR10					
		Untargeted				*Targeted*				*Untargeted*			*Targeted*		
	Model	M1	M2 [6]	M3[4]	M4[7]	M1	M2 [6]	M3 [4]	M4[7]	C1[6]	C2[5]	C3[4]	C1[6]	C2[5]	C3[4]
		Dataset-level Hyperparameter Tuning													
ℓ_0[1]	BB	12	152	52	145	20	179	39	183	28	44	32	29	65	33
	Ours	**9**	**21**	**18**	**15**	**16**	**41**	**28**	**55**	**11**	**17**	**16**	**25**	**38**	**32**
ℓ_1[3]	FAB	8.662	225.728	163.879	312.314	-	-	-	-	-	-	20.475	-	-	-
[1]	BB	10.604	49.834	17.570	46.994	16.602	53.114	29.885	54.312	7.017	10.199	17.134	11.412	15.265	23.367
	Ours	**7.125**	**3.535**	**13.656**	**4.989**	**13.176**	**6.590**	**21.371**	**12.156**	**4.280**	**4.821**	**9.516**	**8.506**	**10.405**	**17.316**
ℓ_2[3]	FAB	1.540	1.591	2.808	16.303	-	-	-	-	0.773	1.113	1.061	-	-	-
[2]	CW	1.633	5.151	3.706	-	2.502	-	4.719	-	0.865	0.997	0.987	1.360	2.900	1.554
[1]	BB	1.747	1.818	3.016	4.571	2.638	2.589	3.519	5.313	0.856	0.950	1.097	1.252	1.455	1.732
[4]	DDN	**1.471**	2.013	2.616	**1.147**	2.308	2.717	3.363	**1.960**	**0.658**	0.771	0.910	1.113	1.307	1.398
	Ours	1.615	**1.089**	**2.607**	1.563	**2.302**	**1.555**	**3.244**	2.407	0.672	**0.741**	**0.910**	**1.091**	**1.281**	**1.380**
ℓ_∞[3]	FAB	0.148	0.365	0.248	0.900	-	-	-	-	0.038	0.052	0.029	-	-	-
[1]	BB	0.159	0.336	0.243	0.409	0.223	0.361	0.280	0.477	0.044	0.054	0.029	0.059	0.074	0.042
	Ours	**0.140**	**0.278**	**0.233**	**0.408**	**0.206**	**0.326**	**0.277**	**0.434**	**0.034**	**0.042**	**0.024**	**0.057**	**0.066**	**0.037**

ℓ_1 norm adversarial examples on MNIST challenge model





Useful links and implementations

Preprint: https://arxiv.org/abs/2102.12827

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

Key results

FMN combines desirable traits of minimum-norm attacks to help improve current adversarial evaluations by:

- finding smaller or comparable minimum-norm perturbations in different ℓ_p norms;
- being less sensitive to hyperparameter choices; and
- being extremely fast by converging quickly and by performing lightweight steps.

We provide extensive experiments and the open source implementation of the attack.

Future work

Extension towards minimum-norm adaptive evaluations.
Improvements that have been suggested for PGD, such as momentum, cyclical step sizes and restarts.
Improvements that overcome obfuscated gradients (e.g. gradient smoothing)

References

[1] W. Brendel *et al.* **Accurate, reliable and fast robustness evaluation**. NeurIPS 2019.
[2] N. Carlini and D. Wagner. **Towards evaluating the robustness of neural networks**. S&P 2017.
[3] F. Croce and M. Hein. **Minimally distorted adversarial examples with a fast adaptive boundary attack**. ICML 2020.
[4] J. Rony *et al.* **Decoupling direction and norm for efficient gradient-based L_2 adversarial attacks and defenses**. CVPR 2019.
[5] Y. Carmon *et al.* **Unlabeled data improves adversarial robustness**. NeurIPS 2019.
[6] A. Madry *et al.* **Towards deep learning models resistant to adversarial attacks**. ICLR 2018.
[7] H. Zhang *et al.* **Towards stable and efficient training of verifiably robust neural networks**. ICLR 2020.

Department of Electrical and Electronic Engineering
University of Cagliari

Piazza d'Armi - 09123 Cagliari, Italy
https://pralab.diee.unica.it