Decoupling Direction and Norm for Efficient Gradient-Based L_2 Adversarial Attacks and Defenses

Jerome Rony, Luiz G. Hafemann, Luiz S. Oliveira, Ismail Ben Ayed, Robert Sabourin, Eric Granger

Laboratoire d'imagerie, de vision et d'intelligence artificielle (LIVIA), ÉTS Montreal, Canada Department of Informatics, Federal University of Paraná, Curitiba, Brazil

Raja Babu Meena(202116010)
Rohan Baghel(202116011)
Group - 14

Dhirubhai Ambani Institute of Information and Communication Technology (DA-IICT), Gandhinagar, Gujarat

July 21, 2025

Overview

Introduction

- To formalize the problem of adversarial examples, the threat model ,and review the main attack and defense method proposed in the literature.
- Objective :
  - low L_2 Norm
  - Miss-classification [1] of the images.

[1] B. Biggio and F. Roli. Wild patterns: Ten years after the rise of adversarial machine learning. Pattern Recognition, 84:317–331, Dec. 2018

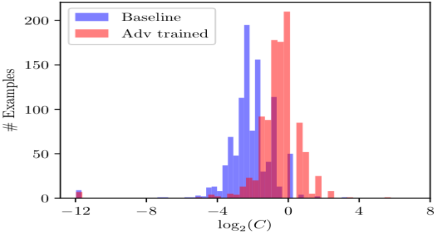
Problem Statement

- Find the smallest perturbation causing miss-classification

$$min_{\delta}||\delta|| \qquad \text{subject to} \qquad argmax\mathbf{P}(y_j|x+\delta,\theta) \neq y_{true}$$
$$\text{and} \qquad 0 \leq x+\delta \leq M$$

Problem Statement

- Problem of C & W [2] L_2 Attack
- $min_\delta ||\delta|| + Cf(x+\delta)$



- Optimal C value is impossible to get for every example
- Changes for adversarially trained models

[2] N. Carlini and D. Wagner. Towards evaluating the robustness of neural networks. In IEEE Symposium on Security and Privacy (SP), pages 39–57, 2017.

Motivation

- Small changes to an image can include miss classification [3].
- Security concern for computer vision applications.

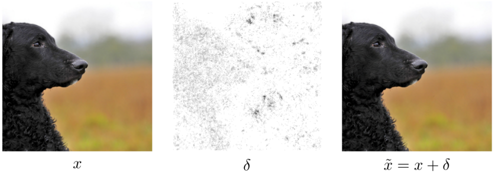


Figure: ImageNet dataset

- The sample x is recognized as a Curly-coated retriever. Adding a perturbation we obtain an adversarial image that is classified as a microwave (with $||\delta||_2 = 0.7$).

[3] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus. Intriguing properties of neural networks. In International Conference on Learning Representations, 2014.

Key assumptions made

- It assumes that there is minimal number of iteration is made (Approx 100 iteration).
- If overfits, overfitting can be reduced easily by L2 Norms.

Approach to solve the problem

- Gradient Based Attack (Decoupled Direction Norm (DDN))
- Instead of imposing a penalty [4], constrain the Norm with a projection.
- In each step, changing the Norm is a binary decision, based on whether the current example in adversarial.

[4] P. A. Jensen and J. F. a. Bard. Operations Research Models and Methods. Wiley, 2003.

Approach to solve the problem

Algorithm 1 Decoupled Direction and Norm Attack

Input: x: original image to be attacked
Input: y: true label (untargeted) or target label (targeted)
Input: K: number of iterations
Input: α: step size
Input: γ: factor to modify the norm in each iteration
Output: $\tilde{x}$: adversarial image

Initialize $\delta_0 \leftarrow \mathbf{0}$, $\tilde{x}_0 \leftarrow x$, $\epsilon_0 \leftarrow 1$
If targeted attack: $m \leftarrow -1$ else $m \leftarrow +1$
for $k \leftarrow 1$ to K **do**
  $g \leftarrow m\nabla_{\tilde{x}_{k-1}} J(\tilde{x}_{k-1}, y, \theta)$
  $g \leftarrow \alpha \frac{g}{\|g\|_2}$ ▷ Step of size α in the direction of g
  $\delta_k \leftarrow \delta_{k-1} + g$
  if $\tilde{x}_{k-1}$ is adversarial **then**
    $\epsilon_k \leftarrow (1-\gamma)\epsilon_{k-1}$ ▷ Decrease norm
  else
    $\epsilon_k \leftarrow (1+\gamma)\epsilon_{k-1}$ ▷ Increase norm
  end if
  $\tilde{x}_k \leftarrow x + \epsilon_k \frac{\delta_k}{\|\delta_k\|_2}$ ▷ Project δ_k onto an ϵ_k-sphere around x
  $\tilde{x}_k \leftarrow \text{clip}(\tilde{x}_k, 0, 1)$ ▷ Ensure $\tilde{x}_k \in \mathcal{X}$
end for
Return $\tilde{x}_k$ that has lowest norm $\|\tilde{x}_k - x\|_2$ and is adversarial

Approach to solve the problem

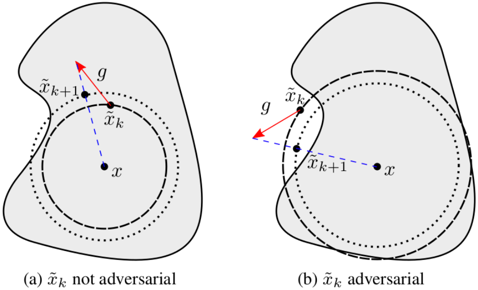


Figure: Illustration of an untargeted attack

- The shaded area denotes the region of the input space classified as y true .

Experimental results for Attack

Dataset	Attack	Budget	Success	Mean L_2	Median L_2	#Grads	Run-time (s)
MNIST	C&W	4×25	100.0	1.7382	1.7400	100	1.7
		1×100	99.4	1.5917	1.6405	100	1.7
		9×10 000	100.0	**1.3961**	1.4121	54 007	856.8
	DeepFool	100	75.4	1.9685	2.2909	98	-
	DDN	100	100.0	1.4563	1.4506	100	1.5
		300	100.0	1.4357	1.4386	300	4.5
		1 000	100.0	1.4240	1.4342	1 000	14.9
CIFAR-10	C&W	4×25	100.0	0.1924	0.1541	60	3.0
		1×100	99.8	0.1728	0.1620	91	4.6
		9×10 000	100.0	0.1543	0.1453	36 009	1 793.2
	DeepFool	100	99.7	0.1796	0.1497	25	-
	DDN	100	100.0	0.1503	0.1333	100	4.7
		300	100.0	0.1487	0.1322	300	14.2
		1 000	100.0	**0.1480**	0.1317	1 000	47.6
ImageNet	C&W	4×25	100.0	1.5812	1.3382	63	379.3
		1×100	100.0	0.9858	0.9587	48	287.1
		9×10 000	100.0	0.4692	0.3980	21 309	127 755.6
	DeepFool	100	98.5	0.3800	0.2655	41	-
	DDN	100	99.6	0.3831	0.3227	100	593.6
		300	100.0	0.3749	0.3210	300	1 779.4
		1 000	100.0	**0.3617**	0.3188	1 000	5 933.6

- Performance of our DDN attack compared to C & W and DeepFool [5] attacks on MNIST, CIFAR-10 and ImageNet in the untargeted scenario.

[5] S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard. Deepfool: a simple and accurate method to fool deep neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 2574–2582, 2016.

Experimental results for Defense

Defense evaluation

Dataset	Defense	Mean L_2	Accuracy at $\|\delta\| \leq \epsilon$
MNIST $\epsilon = 1.5$	Baseline	1.3778	40.8
	Madry	1.6917	67.3
	Ours	**2.4497**	**87.2**
CIFAR-10 $\epsilon = 0.5$	Baseline	0.1282	0.1
	Madry	0.6601	56.1
	Ours	**0.8597**	**67.6**

Higher Mean L_2 is better

Conclusions

- DDN obtains comparable results with the state-of-the-art for L_2 norm adversarial perturbations, but in much fewer iterations.
- Attack allows for faster evaluation of the robustness of differentiable models, and enables a novel adversarial training.
- Our experiments with MNIST and CIFAR-10 show state-of-the-art robustness against L_2 -based attacks in a white-box scenario.

Many Thanks

- This research was supported by the Fonds de recherche du Québec - Nature et technologies, Nat- ural Sciences and Engineering Research Council of Canada, and CNPq grant 206318/2014-6.