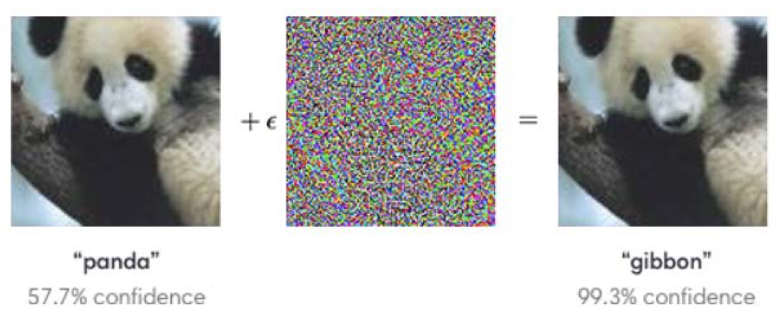
Reinforced Adversarial Attacks on Deep Neural Networks Using ADMM

Pu Zhao[1], Kaidi Xu[1], Tianyun Zhang[2], Makan Fardad[2], Yanzhi Wang[1], Xue Lin[1]
[1]Department of ECE, Northeastern University;
[2]College of Engineering & Computer Science, Syracuse University;

Introduction

- Deep neural networks (DNNs) are known vulnerable to adversarial attacks
- Adversarial examples in adversarial attacks:
  adding delicately crafted distortions onto original legal inputs, can mislead a DNN to classify them as any target labels.
- L_p norms of the distortion:
  the added distortions are usually measured by $L_0, L_1, L_2,$ and L_∞ norms in $L_0, L_1, L_2,$ and L_∞ attacks.



- A unified framework:
  this work for the first time unifies the methods of generating adversarial examples by leveraging ADMM. $L_0, L_1, L_2,$ and L_∞ attacks are effectively implemented by this general framework with little modifications.

Notations and Definitions

- Representations of the DNN model:
  input: $x \in R^{hw}$ or $x \in R^{3hw}$
  model: $F(x) = y$
  output: $0 \le y_i \le 1$ and $y_1 + y_2 + \cdots + y_m = 1$
  logits: $F(x) = \text{softmax}(Z(x)) = y$
  classification: $C(x) = \arg\max y_i$
  distance:

$$\|x - x_0\|_p = \left(\sum_{i=1}^{n} |x_i - x_{0i}|^p\right)^{\frac{1}{p}}$$

- Adversarial attack:

$$\underset{\delta}{\text{minimize}} \quad D(\delta) + g(x + \delta)$$
$$\text{subject to} \quad (x + \delta) \in [0, 1]^n,$$

$Z(x)$: logits before softmax layer

$$g(x) = c \cdot \max\left(\left(\max_{i \ne t}(Z(x)_i) - Z(x)_t\right), -\kappa\right)$$

ADMM Formulation

- Reformulate the original problem:

$$\underset{\delta,\mathbf{z},\mathbf{w}}{\text{minimize}} \quad D(\delta) + g(\mathbf{x} + \mathbf{z}) + h(\mathbf{w})$$
$$\text{subject to} \quad \mathbf{z} = \delta$$
$$\mathbf{w} = \mathbf{x} + \mathbf{z},$$

$$h(\mathbf{w}) = \begin{cases} 0 & \mathbf{w} \in [0,1]^n \\ \infty & \text{otherwise.} \end{cases}$$

- The augmented Lagrangian function:

$$L(\delta, \mathbf{z}, \mathbf{w}, \mathbf{u}, \mathbf{v}) = D(\delta) + g(\mathbf{x} + \mathbf{z}) + h(\mathbf{w}) + \mathbf{u}^T(\delta - \mathbf{z}) + \mathbf{v}^T(\mathbf{w} - \mathbf{z} - \mathbf{x}) + \frac{\rho}{2}\|\delta - \mathbf{z}\|_2^2 + \frac{\rho}{2}\|\mathbf{w} - \mathbf{z} - \mathbf{x}\|_2^2,$$

General Framework based on ADMM

- ADMM iterations
  In the k-th iteration, the following steps are performed:

$$\{\delta^{k+1}, \mathbf{w}^{k+1}\} = \arg\min L(\delta, \mathbf{z}^k, \mathbf{w}, \mathbf{u}^k, \mathbf{v}^k)$$
$$\mathbf{z}^{k+1} = \arg\min L(\delta^{k+1}, \mathbf{z}, \mathbf{w}^{k+1}, \mathbf{u}^k, \mathbf{v}^k)$$
$$\mathbf{u}^{k+1} = \mathbf{u}^k + \rho(\delta^{k+1} - \mathbf{z}^{k+1})$$
$$\mathbf{v}^{k+1} = \mathbf{v}^k + \rho(\mathbf{w}^{k+1} - \mathbf{x}^{k+1} - \mathbf{z}^{k+1}).$$

$$\underset{\delta}{\text{minimize}} \quad D(\delta) + \frac{\rho}{2}\|\delta - \mathbf{z}^k + (1/\rho)\mathbf{u}^k\|_2^2$$
$$\underset{\mathbf{w}}{\text{minimize}} \quad h(\mathbf{w}) + \frac{\rho}{2}\|\mathbf{w} - \mathbf{z}^k - \mathbf{x} + (1/\rho)\mathbf{v}^k\|_2^2.$$
$$\underset{\mathbf{z}}{\text{minimize}} \quad g(\mathbf{x} + \mathbf{z}) + \frac{\rho}{2}\|\delta^{k+1} - \mathbf{z} + (1/\rho)\mathbf{u}^k\|_2^2 + \frac{\rho}{2}\|\mathbf{w}^{k+1} - \mathbf{z} - \mathbf{x} + (1/\rho)\mathbf{v}^k\|_2^2,$$

- w step

$$\underset{\mathbf{w}}{\text{minimize}} \quad h(\mathbf{w}) + \frac{\rho}{2}\|\mathbf{w} - \mathbf{z}^k - \mathbf{x} + (1/\rho)\mathbf{v}^k\|_2^2.$$

$$[\mathbf{w}^{k+1}]_i = \begin{cases} 0 & \text{if } [\mathbf{z}^k + \mathbf{x} - (1/\rho)\mathbf{v}^k]_i < 0 \\ 1 & \text{if } [\mathbf{z}^k + \mathbf{x} - (1/\rho)\mathbf{v}^k]_i > 1 \\ [\mathbf{z}^k + \mathbf{x} - (1/\rho)\mathbf{v}^k]_i & \text{otherwise,} \end{cases}$$

- z step

$$\underset{\mathbf{z}}{\text{minimize}} \quad g(\mathbf{x} + \mathbf{z}) + \frac{\rho}{2}\|\delta^{k+1} - \mathbf{z} + (1/\rho)\mathbf{u}^k\|_2^2 + \frac{\rho}{2}\|\mathbf{w}^{k+1} - \mathbf{z} - \mathbf{x} + (1/\rho)\mathbf{v}^k\|_2^2,$$

$$\underset{\mathbf{z}}{\text{minimize}} \quad (\nabla g(\mathbf{z}^k + \mathbf{x}))^T(\mathbf{z} - \mathbf{z}^k) + \frac{1}{2}\|\mathbf{z} - \mathbf{z}^k\|_{\mathbf{G}}^2 + \frac{\rho}{2}\|\mathbf{z} - \mathbf{a}\|_2^2 + \frac{\rho}{2}\|\mathbf{z} - \mathbf{b}\|_2^2.$$

$$\mathbf{z}^{k+1} = \frac{1}{\alpha + 2\rho}(\alpha\mathbf{z}^k + \rho\mathbf{a} + \rho\mathbf{b} - \nabla g(\mathbf{z}^k + \mathbf{x}))$$

$\nabla g(\mathbf{z}^k + \mathbf{x})$
first-order Taylor expansion | Bregman divergence
$(\nabla g(\mathbf{z}^k + \mathbf{x}))^T(\mathbf{z} - \mathbf{z}^k) + \frac{1}{2}\|\mathbf{z} - \mathbf{z}^k\|_{\mathbf{G}}^2$

Four Attacks based on the Framework

- Proximal operator

$$\mathbf{prox}_{\lambda D}(s) = \arg\min_{\delta}\left(\lambda D(\delta) + \frac{1}{2}\|\delta - s\|_2^2\right)$$

- L2 attack:

$$\mathbf{prox}_{\lambda 2}(\mathbf{s}) = \arg\min_{\delta}\left(\lambda\|\delta\|_2 + \frac{1}{2}\|\delta - \mathbf{s}\|_2^2\right) \Rightarrow \mathbf{prox}_{\lambda 2}(\mathbf{s}) = \begin{cases} (1 - \lambda/\|\mathbf{s}\|_2)\mathbf{s} & \|\mathbf{s}\|_2 \ge \lambda \\ 0 & \|\mathbf{s}\|_2 < \lambda \end{cases}$$

- L0 attack

$$\mathbf{prox}_{\lambda 0}(\mathbf{s}) = \arg\min_{\delta}\left(\lambda\|\delta\|_0 + \frac{1}{2}\|\delta - \mathbf{s}\|_2^2\right) \Rightarrow (\mathbf{prox}_{\lambda 0}(\mathbf{s}))_i = \begin{cases} 0 & |s_i| < \sqrt{2\lambda} \\ 0 \text{ or } s_i & |s_i| = \sqrt{2\lambda} \\ s_i & |s_i| > \sqrt{2\lambda} \end{cases}$$

- L1 attack

$$\mathbf{prox}_{\lambda 1}(\mathbf{s}) = \arg\min_{\delta}\left(\lambda\|\delta\|_1 + \frac{1}{2}\|\delta - \mathbf{s}\|_2^2\right) \Rightarrow (\mathbf{prox}_{\lambda 1}(\mathbf{s}))_i = \begin{cases} s_i - \lambda & s_i \ge \lambda \\ 0 & |s_i| < \lambda \\ s_i + \lambda & s_i \le -\lambda \end{cases}$$

- L infinity attack

$$\underset{\delta}{\text{minimize}} \quad \|\delta\|_\infty + \frac{\rho}{2}\|\delta - \mathbf{s}\|_2^2,$$

It has no closed form solution. We can obtain its solution by derive its KKT condition.

$$\sum_{i=1}^{n} \rho(s_i - t^*)_+ = 1 \qquad \delta_i^* = \min\{t^*, s_i\}$$

Experimental Results

- Adversarial attacks:

L_0 attack

Dataset	Attack method	Best case		Average case		Worst case	
		ASR	L_0	ASR	L_0	ASR	L_0
MNIST	C&W(L_0)	100	7.88	100	16.58	100	29.84
	ADMM(L_0)	100	6.94	100	13.35	100	23.66
CIFAR	C&W(L_0)	100	8.16	100	20.82	100	35.07
	ADMM(L_0)	100	7.64	100	18.78	100	32.81

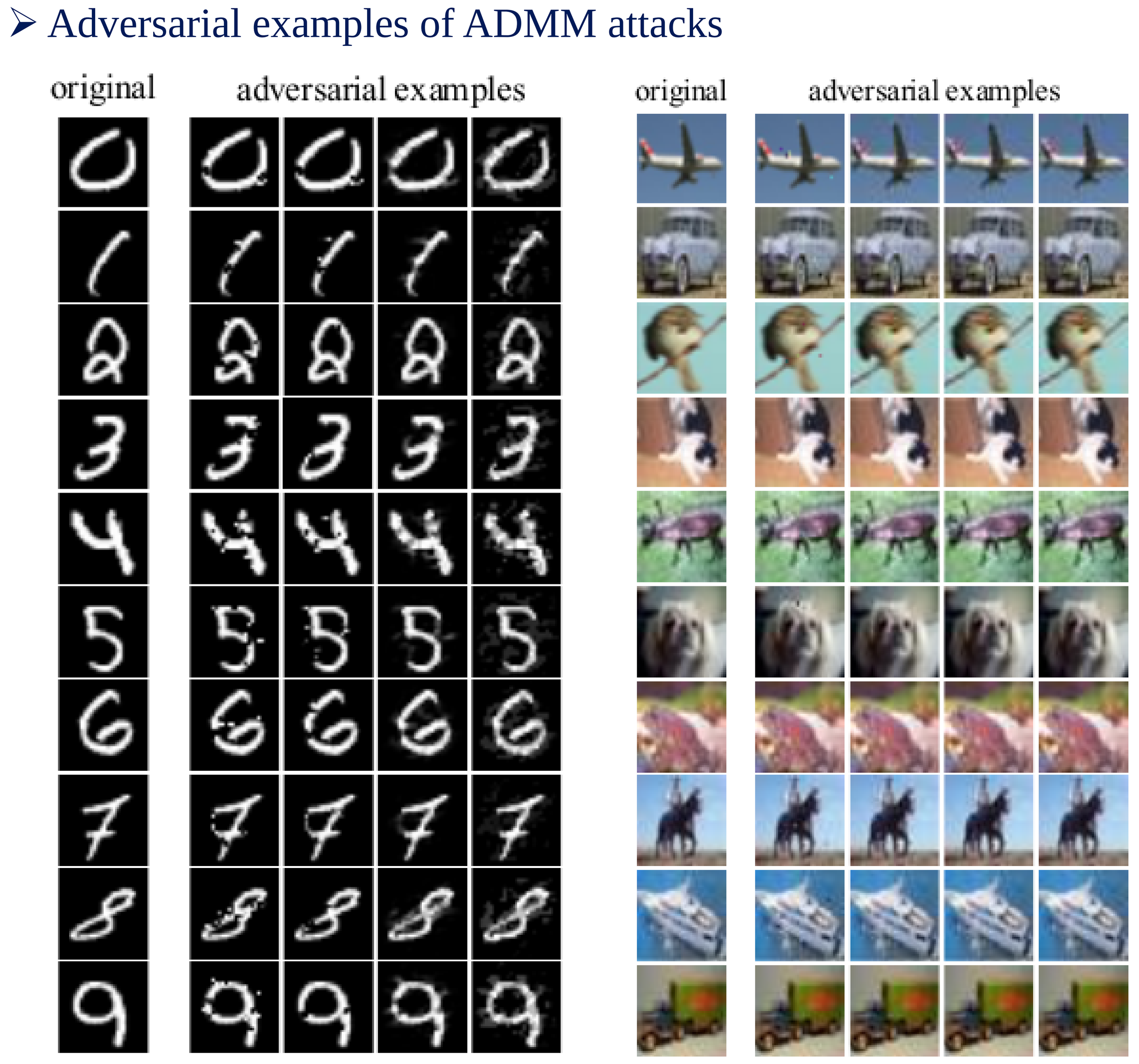
L_1 attack

Data Set	Methods	Best Case		Average Case		Worst Case	
		ASR	L_1	ASR	L_1	ASR	L_1
MNIST	IFGM(L_1)	100	17.3	100	34.6	100	58.4
	EAD(L_1)	100	7.74	100	14.16	100	21.38
	ADMM(L_1)	100	6.29	100	12.35	100	17.9
CIFAR-10	IFGM(L_1)	100	5.96	100	15.8	100	20.8
	EAD(L_1)	100	1.94	100	4.62	100	7.25
	ADMM(L_1)	100	1.75	100	3.750	100	5.92
ImageNet	IFGM(L_1)	100	298	100	580	100	685
	EAD(L_1)	100	60.98	100	112.7	100	185
	ADMM(L_1)	100	49.17	100	75.2	100	127

L_2 attack

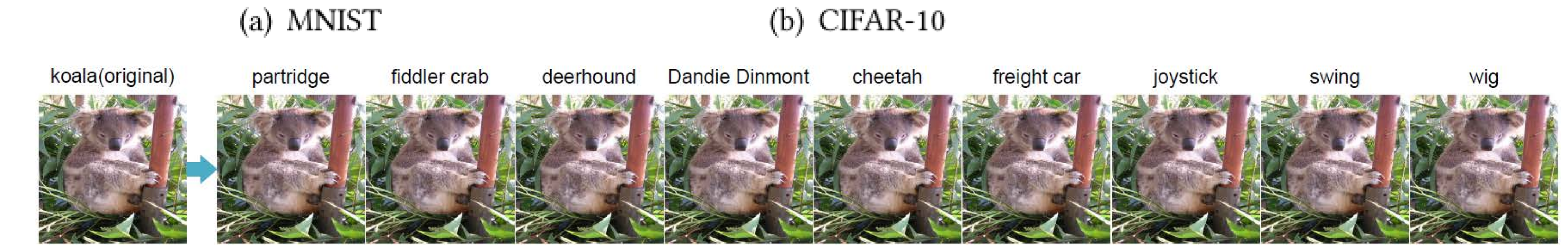
Data Set	Attack Method	Best Case				Average Case				Worst Case			
		ASR	L_2	L_1	L_∞	ASR	L_2	L_1	L_∞	ASR	L_2	L_1	L_∞
MNIST	FGM(L_2)	99.3	2.158	23.7	0.562	43.2	3.18	37.6	0.761	0	N.A.	N.A.	N.A.
	IFGM(L_2)	100	1.61	18.2	0.393	99.7	2.43	31.8	0.574	99.3	3.856	54.1	0.742
	C&W(L_2)	100	1.356	13.32	0.394	100	1.9	21.11	0.533	99.6	2.52	30.44	0.673
	ADMM(L_2)	100	1.268	15.93	0.398	100	1.779	25.06	0.444	99.9	2.269	34.7	0.561
CIFAR-10	FGM(L_2)	99.7	0.418	13.85	0.05	40.6	1.09	37.4	0.62	1.2	4.17	119.3	0.43
	IFGM(L_2)	100	0.185	6.26	0.021	100	0.419	14.9	0.043	100	0.685	22.8	0.0674
	C&W(L_2)	100	0.170	5.721	0.0189	100	0.322	11.28	0.0347	100	0.445	15.79	0.0495
	ADMM(L_2)	100	0.163	5.66	0.0192	100	0.315	10.97	0.0354	100	0.427	15.05	0.0502
ImageNet	FGM(L_2)	15	2.37	815	0.129	3	7.51	2104	0.25	0	N.A.	N.A.	N.A.
	IFGM(L_2)	100	0.984	328	0.031	100	2.38	795	0.079	97.6	4.59	1354	0.177
	C&W(L_2)	100	0.449	126.8	0.0159	100	0.621	198	0.0218	100	0.81	272.3	0.031
	ADMM(L_2)	100	0.412	112.5	0.017	100	0.555	166.7	0.021	100	0.704	225.6	0.0356

- Adversarial examples of ADMM attacks



(a) MNIST (b) CIFAR-10

Original input has label t. The adversarial examples can mislead the DNN to classify them as label t+2.



Adversarial examples on ImageNet, where an input of koala can be classified as other target labels by adding small distortions.