Structured Adversarial Attack Against Deep Neural Networks

Kaidi Xu[1,2] (Joint work with Sijia Liu[3], Xue Lin[1]), Mentor: Bhavya Kailkhura[2]
1. Northeastern University 2. DSSI, LLNL 3. MIT-IBM Research

DATA SCIENCE SUMMER INSTITUTE

Lawrence Livermore National Laboratory

Introduction

Neural network classifiers are easily fooled by adversarial perturbations

original image

Africa elephant

add imperceptible perturbation

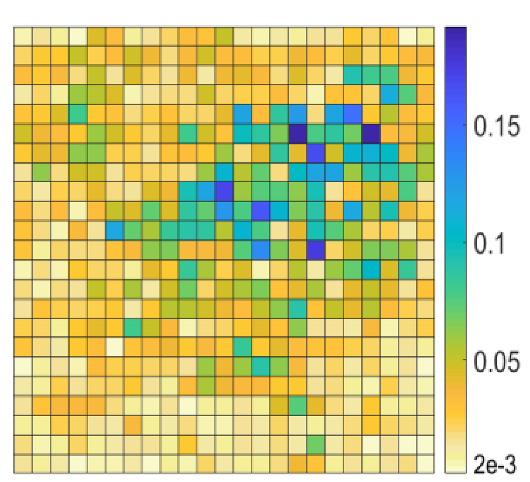
adversarial example

Dinning table

Motivations

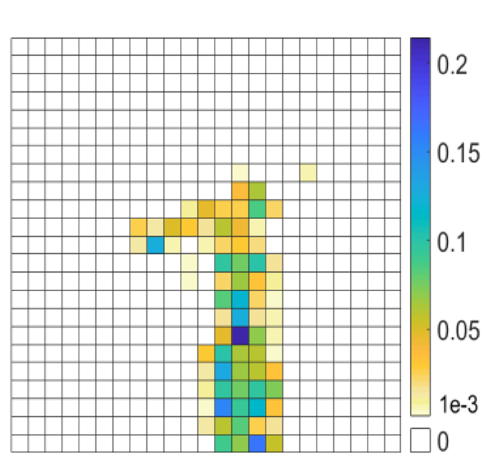
★ **Is it effective to add perturbation on whole image?**

attack original image 'ostrich' to 'unicycle'



perturbation heatmap generated by C&W[2] attack

★ **If not, how can we capture most effective area to attack?**



perturbation heatmap generated by StrAttack attack

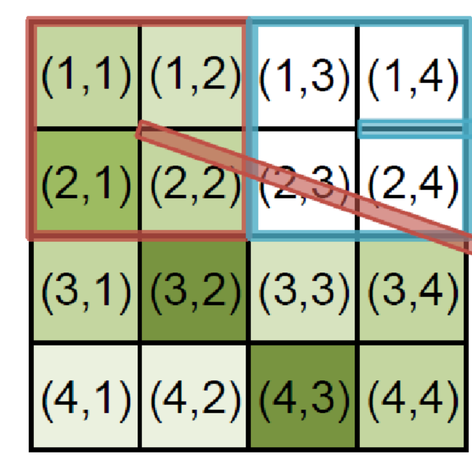
Meet interpretability!

Keep same level Lp norms with other attack methods; see experiment results

Our General Gramework: StrAttack

1) group sparsity

non-overlapping groups



•L2 norm penalizes the group of pixels as a unit

$$\|\Delta_{\mathcal{G}_{1,2}}\|_2 = \sqrt{\sum_{(i,j)\in\mathcal{G}_{1,2}} \Delta_{ij}^2}$$

$$\|\Delta_{\mathcal{G}_{1,1}}\|_2 = \sqrt{\sum_{(i,j)\in\mathcal{G}_{1,1}} \Delta_{ij}^2}$$

2) Optimization problem:

attack loss

overall ℓ_1 or ℓ_2 distortion $D(\boldsymbol{\delta}) = |\boldsymbol{\delta}|_1$

group sparsity $g(\boldsymbol{\delta}) = \sum_i \|\boldsymbol{\delta}_{\mathcal{G}_i}\|_2$

$$\begin{array}{ll} \underset{\boldsymbol{\delta}}{\text{minimize}} & f(\mathbf{x}_0+\boldsymbol{\delta}, t) + \gamma D(\boldsymbol{\delta}) + \tau g(\boldsymbol{\delta}) \\ \text{subject to} & (\mathbf{x}_0+\boldsymbol{\delta}) \in [0,1]^n,\ \|\boldsymbol{\delta}\|_\infty \le \epsilon, \end{array}$$

bounded box

pixel-level distortion

•**Challenges:**
- Smooth + nonsmooth composite optimization
- Multiple constraints, hard & soft

3) Solve by ADMM

$$\begin{array}{ll} \text{minimize} & f(\boldsymbol{\delta}) + \gamma D(\mathbf{z}) + \tau g(\mathbf{y}) + I_C(\mathbf{w}) \\ \text{subject to} & \boldsymbol{\delta} = \mathbf{z}, \quad \boldsymbol{\delta} = \mathbf{y}, \quad \boldsymbol{\delta} = \mathbf{w} \end{array}$$

Lagrangian multipliers (dual variables)

$$L(\boldsymbol{\delta}, \mathbf{z}, \mathbf{y}, \mathbf{w}, \mathbf{u}, \mathbf{v}, \mathbf{s}) = f(\boldsymbol{\delta}) + \gamma D(\mathbf{z}) + \tau g(\mathbf{y}) + I_C(\mathbf{w}) + \mathbf{u}^T(\boldsymbol{\delta} - \mathbf{z}) + \mathbf{v}^T(\boldsymbol{\delta} - \mathbf{y}) + \mathbf{s}^T(\boldsymbol{\delta} - \mathbf{w}) + \frac{\rho}{2}\|\boldsymbol{\delta} - \mathbf{z}\|_2^2 + \frac{\rho}{2}\|\boldsymbol{\delta} - \mathbf{y}\|_2^2 + \frac{\rho}{2}\|\boldsymbol{\delta} - \mathbf{w}\|_2^2$$

Alternate between ***Red*** & ***Blue***

$$\begin{cases} \boldsymbol{\delta}_{t+1} = \operatorname{argmin}_{\boldsymbol{\delta}} \{L(\boldsymbol{\delta}, \mathbf{z}_t, \mathbf{y}_t, \mathbf{w}_t, \mathbf{u}_t, \mathbf{v}_t, \mathbf{s}_t)\} \\ \mathbf{z}_{t+1}, \mathbf{y}_{t+1}, \mathbf{w}_{t+1} = \operatorname{argmin}_{\mathbf{z},\mathbf{y},\mathbf{w}} \{L(\boldsymbol{\delta}_{t+1}, \mathbf{z}, \mathbf{y}, \mathbf{w}, \mathbf{u}_t, \mathbf{v}_t, \mathbf{s}_t)\} \\ \mathbf{u}_{t+1} = \mathbf{u}_t - \rho(\boldsymbol{\delta}_{t+1} - \mathbf{z}_{t+1}), \quad \mathbf{v}_{t+1} = \mathbf{v}_t - \rho(\boldsymbol{\delta}_{t+1} - \mathbf{y}_{t+1}), \quad \mathbf{s}_{t+1} = \mathbf{s}_t - \rho(\boldsymbol{\delta}_{t+1} - \mathbf{w}_{t+1}) \end{cases}$$

First-order (or **zeroth-order** if *f* is black box)

Decomposed in ***z***, ***y***, ***w*** ***(closed-form)***

Dual updates

Experimental Results

• **Attacking performance**

Table 1: Adversarial attack success rate (ASR) and ℓ_p distortion values for various attacks.

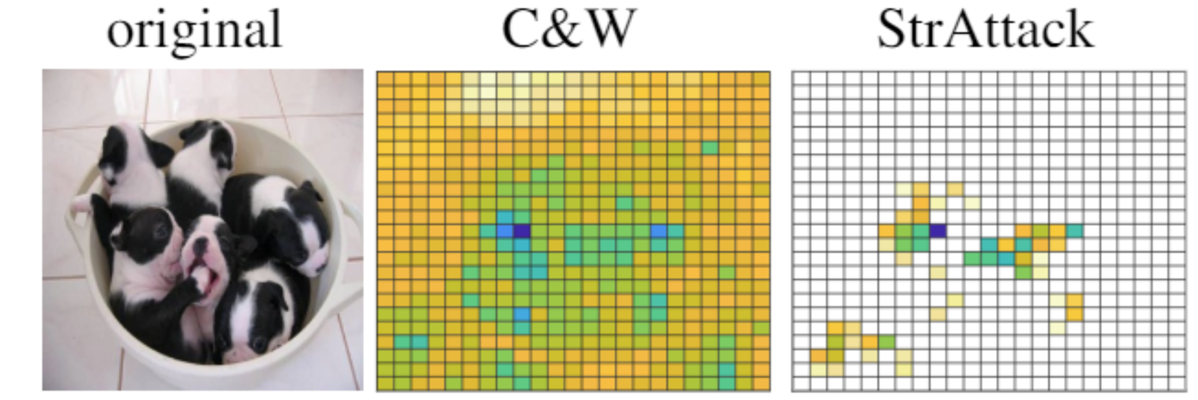
Data Set	Attack Method	Best Case*					Average Case*					Worst Case*				
		ASR	ℓ_0	ℓ_1	ℓ_2	ℓ_∞	ASR	ℓ_0	ℓ_1	ℓ_2	ℓ_∞	ASR	ℓ_0	ℓ_1	ℓ_2	ℓ_∞
MNIST	FGM	99.3	456.5	28.2	2.32	0.57	35.8	466	39.4	3.17	0.717	0	N.A.**	N.A.	N.A.	N.A.
	IFGSM	100	549.5	18.3	1.57	0.4	100	588	30.9	2.41	0.566	99.8	640.4	50.98	3.742	0.784
	C&W	100	479.8	13.3	1.35	0.397	100	493.4	21.3	**1.9**	0.528	99.7	524.3	29.9	**2.45**	0.664
	StrAttack	100	**73.2**	10.9	1.51	**0.384**	100	**119.4**	18.05	2.16	**0.47**	100	**182.0**	26.9	2.81	**0.5**
	+overlap	100	84.4	**9.2**	**1.32**	0.401	100	157.4	**16.2**	1.95	0.508	100	260.9	**22.9**	2.501	0.653
CIFAR-10	FGM	98.5	3049	12.9	0.389	0.046	44.1	3048	34.2	0.989	0.113	0.2	3071	61.3	1.76	0.194
	IFGSM	100	3051	6.22	0.182	0.02	100	3051	13.7	0.391	0.0433	100	3060	22.9	0.655	0.075
	C&W	100	2954	6.03	0.178	**0.019**	100	2956	12.1	0.347	**0.0364**	99.9	3070	16.8	**0.481**	**0.0536**
	StrAttack	100	**264**	**3.33**	0.204	0.031	100	**487**	7.13	0.353	0.050	100	**772**	**12.5**	0.563	0.075
	+overlap	100	295	3.35	**0.169**	0.029	100	562	**7.05**	**0.328**	0.047	100	920	12.9	0.502	0.063
ImageNet	FGM	12	264917	152	0.477	**0.0157**	2	263585	51.3	**0.18**	**0.00614**	0	N.A.	N.A.	N.A.	N.A.
	IFGSM	100	267079	299.32	0.9086	0.02964	100	267293	723	2.2	0.0792	98	267581	1378	4.22	0.158
	C&W	100	267916	127	**0.471**	0.016	100	263140	198	0.679	0.03	100	265212	268	**0.852**	**0.041**
	StrAttack	100	**14462**	**55.2**	0.719	0.058	100	**52328**	**152**	1.06	0.075	100	**80722**	**197**	1.35	0.122

* Please refer to Appendix F for the definition of best case, best case and worst case.
** N.A. means not available in the case of zero ASR, +overlap means structured attack with overlapping groups.

- Minimum number of perturbed pixels
- Same-level $\ell_1, \ell_2, \ell_\infty$ perturbation strength

• **Interpretability by CAM**

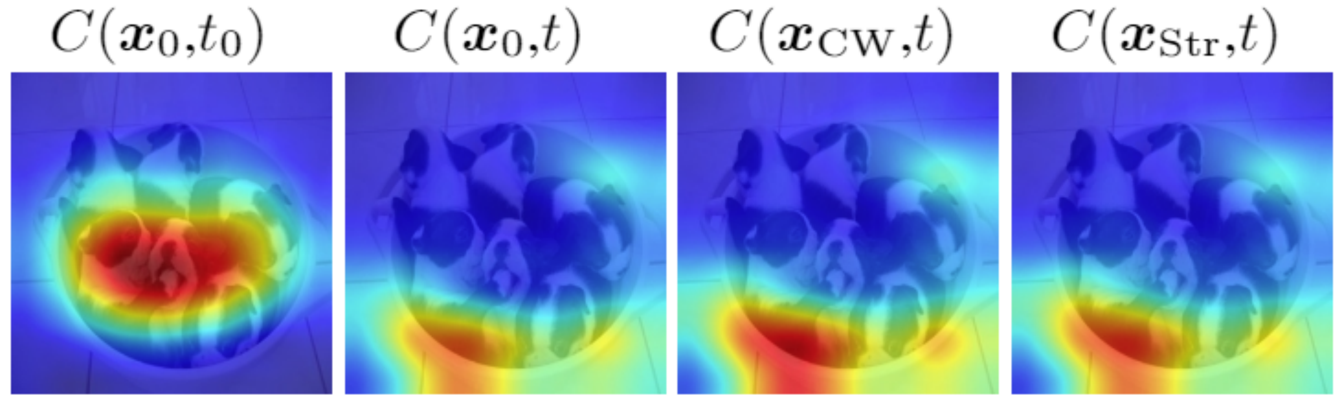
original C&W StrAttack



t_0: Boston bull, t: bucket

Perturbing the area that either Boston bull or the bucket located, which fits CAM visualization.

$C(\boldsymbol{x}_0, t_0)$ $C(\boldsymbol{x}_0, t)$ $C(\boldsymbol{x}_{\mathrm{CW}}, t)$ $C(\boldsymbol{x}_{\mathrm{Str}}, t)$



Current Work

In the future, we will focus on the verification problem (certifying that no small perturbations of a given input can cause the neural network to change its prediction) on any existing compute graph. The research on this topic will lead us a guaranteed robust error and shed the light on the provable robustness of neural networks.

Please keep in mind that deep neural networks are easy to be attacked!

This work performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.

IM release number: LLNL-POST-782239