Group-wise Sparse Adversarial Attacks

S. Sadiku, M. Wagner, S. Pokutta

ZIB ZUSE INSTITUTE BERLIN

Berlin Mathematical School BMS

Motivation

- *Deep Neural Networks (DNN)* are vulnerable to adversarial attacks
- *Sparse adversarial attacks* explore ℓ_p neighborhoods with $p = 0$ via
  1. Greedy single-pixel selection
  2. Local search techniques
  3. Evolutionary algorithms
  4. Relaxing ℓ_0 via the ℓ_1 ball
- Such methods do not constrain the magnitude of changed pixels
- Generate adversarial attacks that are simultaneously sparse and imperceptible
- Impose structure to sparse adversarial attacks by generating group-wise sparse perturbations that are targeted to the main objective in the image - leads to *explainable* perturbations
- Sheds light on significant vulnerabilities in DNNs and offers insights into their failures

Group-wise sparse Attacks

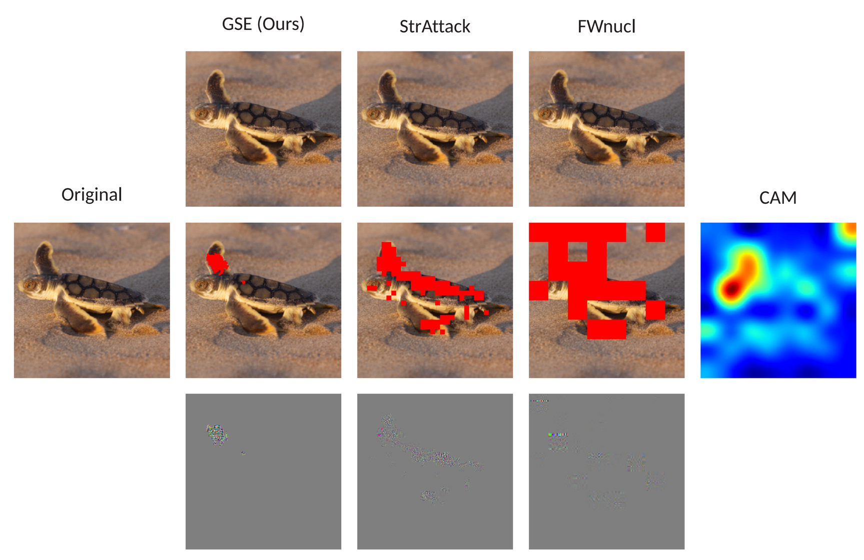

Figure 1: Visual comparison of successful untargeted adversarial instances generated by our attack (1), StrAttack (2), and FWnucl (3). The target model is a ResNet50.

Method

- $\mathscr{X} = [I_{\text{min}}, I_{\text{max}}]^{M\times N\times C}$ set of feasible images
- $\mathscr{L} : \mathscr{X} \times \mathbb{N} \to \mathbb{R}$ classification loss function
- *Targeted sparse adversarial attacks* for given $\mathbf{x}$, target t

$$\min_{\mathbf{w}\in\mathbb{R}^{M\times N\times C}} \mathscr{L}(\mathbf{x}+\mathbf{w}, t) + \lambda\|\mathbf{w}\|_p^p,$$

- Use forward-backward splitting algorithm for $0 < p < 1$
- Requires solving the proximal operator

$$\text{prox}_{\lambda\|\cdot\|_p^p}(\mathbf{w}) := \operatorname*{arg\,min}_{\mathbf{y}\in\mathbb{R}^{M\times N\times C}} \frac{1}{2\lambda}\|\mathbf{y}-\mathbf{w}\|_2^2 + \|\mathbf{y}\|_p^p,$$

- Closed-form solution for $p = 1/2$ and $p = 2/3$
- Tune λ to determine pixel coordinates to perturb
  1. Build a mask $\mathbf{m} = \text{sign}\left(\sum_{c=1}^{C} |\mathbf{w}^{(k)}|_{:,:,c}\right) \in \{0,1\}^{M\times N}$,
  2. Apply Gaussian blur kernel $\mathbf{M} = \mathbf{m} ** \mathbf{K} \in [0,1]^{M\times N}$,
  3. Build

$$\overline{\mathbf{M}}_{ij} = \begin{cases} \mathbf{M}_{ij}+1, & \text{if } \mathbf{M}_{ij} \neq 0, \\ q, & \text{else}, \end{cases}$$

  4. Set $\lambda_{i,j,:}^{(k+1)} = \frac{\lambda_{i,j,:}^{(k)}}{\overline{\mathbf{M}}_{i,j}}$
- Nesterov Accelerated Gradient over chosen coordinates

Speed Comparison

	Untargeted			Targeted		
	CIFAR-10	ImageNet		CIFAR-10	ImageNet	
Attack	ResNet20	ResNet50	VGG19	ResNet20	ResNet50	VGG19
GSE (Ours)	**0.39s**	**23.8s**	**38.9s**	**0.39s**	**20.1s**	**40.8s**
StrAttack	1.33s	48.9s	78.2s	1.28s	49.2s	75.5s
FWnucl	0.80s	32.4s	67.1s	0.82s	31.9s	65.8s

Evaluation metrics and Results on Untargeted Attacks

- $(\mathbf{x}^{(i)})_{0<i\leq n}$ images of perturbation $(\delta^{(i)})_{0<i\leq n}$
- *Attack Success Rate* ASR $= \frac{m}{n}$ for m successful adversaries
- *Average Number of Changed Pixels*

$$\text{ACP} = \frac{1}{n}\sum_{i=1}^{n} \frac{\|\Delta^{(i)}\|_0}{MN}, \quad \Delta^{(i)} = \sum_{c=1}^{C} |\delta_{[:,:,c]}^{(i)}| \in \mathbb{R}^{M\times N}$$

- Run depth-first search (DFS) on $\mathbf{m} = \text{sign}\left(\sum_{c=1}^{C} |\delta_{[:,:,c]}^{(i)}|\right)$ starting from every $1-$entry another DFS has not yet discovered
- *Average Number of Clusters (ANC)* - average the number of DNS runs until all $1-$entries are discovered for m adversaries
- For $n < M, N$ and $\mathscr{G} = \{G_1, ..., G_k\}$ a set containing the index sets of all overlapping n by n patches in δ

$$d_{2,0}(\delta) := |\{i \,:\, \|\delta_{G_i}\|_2 \neq 0,\ i = 1, ..., k\}|$$

	Attack	ASR	ACP	ANC	ℓ_2	$d_{2,0}$
CIFAR-10 ResNet20	GSE (Ours)	**100%**	**36.8**	1.5	**0.75**	**177**
	StrAttack	**100%**	117	4.7	1.07	419
	FWnucl	95.1%	456	**1.3**	2.00	592
NIPS2017 VGG19	GSE (Ours)	**100%**	**968**	6.7	**1.25**	**2596**
	StrAttack	**100%**	4021	7.4	1.92	7058
	FWnucl	89.1%	7225	**2.1**	2.40	7985
NIPS2017 ResNet50	GSE (Ours)	**100%**	**1270**	8.2	**1.47**	**2922**
	StrAttack	**100%**	8669	12.9	2.51	13963
	FWnucl	48.6%	14953	**3.7**	1.82	17083

- ResNet20 classifier for CIFAR-10
- VGG19 and a ResNet50 classifier for ImageNet/NIPS2017
- Tested on 1,000 samples from each dataset

Results on Targeted Attacks

- ResNet20 classifier for CIFAR-10
- VGG19 and a ResNet50 classifier for NIPS2017
- Tested on 1,000 samples from each dataset

		Best case					Average case					Worst case				
Dataset	Attack	ASR	ACP	ANC	ℓ_2	$d_{2,0}$	ASR	ACP	ANC	ℓ_2	$d_{2,0}$	ASR	ACP	ANC	ℓ_2	$d_{2,0}$
CIFAR-10 ResNet20	GSE (Ours)	**100%**	**29.6**	1.2	**0.69**	**151**	**100%**	**80.1**	2.1	**1.17**	**276**	**100%**	**148**	3.5	**1.62**	**413**
	StrAttack	**100%**	75.4	2.2	0.79	336	**100%**	232	5.4	1.96	532	**100%**	430	9.0	4.72	620
	FWnucl	**100%**	276	**1.0**	1.44	504	82.8%	384	**1.6**	2.32	567	35.0%	471	**2.8**	3.78	602
NIPS2017 VGG19	GSE (Ours)	**100%**	**1974**	4.9	**2.72**	**3695**	**100%**	**6025**	10.5	**3.58**	**9704**	**100%**	**15996**	**17.2**	**4.39**	**21489**
	StrAttack	**100%**	3616	3.6	2.85	5863	**100%**	10940	11.0	3.70	16545	**100%**	23245	19.9	5.48	32193
	FWnucl	56.1%	4489	**1.3**	2.91	5783	18.3%	7133	**1.9**	3.92	11639	0.0%	N/A	N/A	N/A	N/A
NIPS2017 ResNet50	GSE (Ours)	**100%**	**2090**	3.8	**2.51**	**3498**	**100%**	**7311**	9.5	**3.15**	**10734**	**100%**	**16284**	**15.9**	**3.65**	**21872**
	StrAttack	**100%**	6117	4.0	2.73	9246	**100%**	15308	12.1	4.17	21182	**100%**	26569	20.5	7.88	33297
	FWnucl	32.4%	9897	**3.4**	2.82	11134	12.6%	11735	**6.3**	3.96	18126	0.0%	N/A	N/A	N/A	N/A

Visual Anylsis

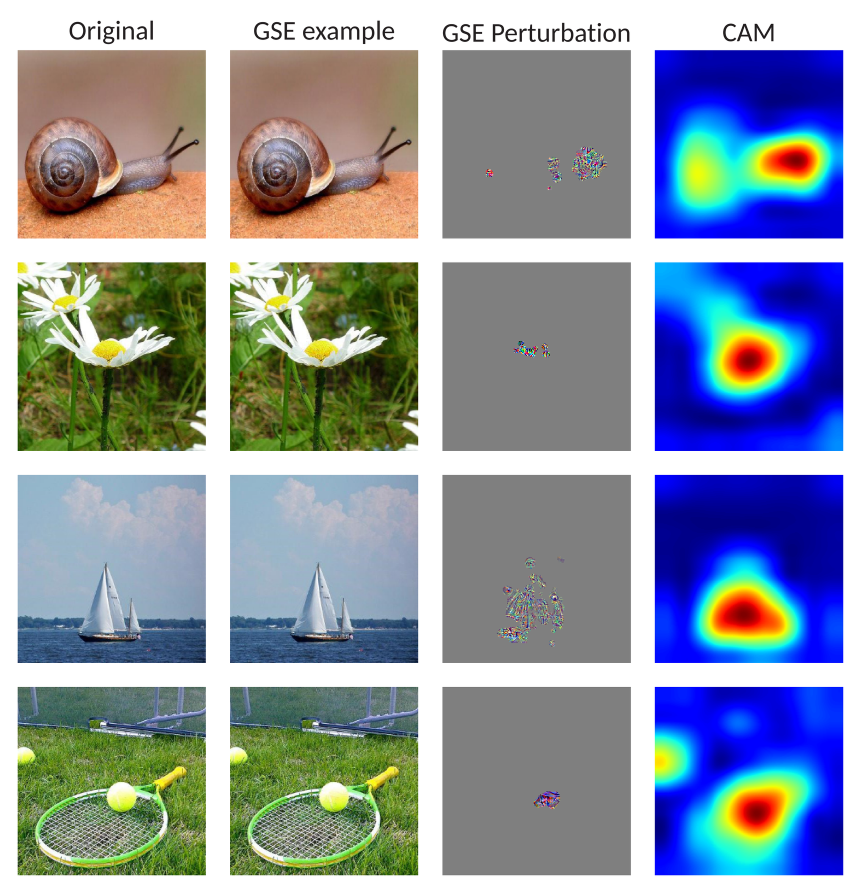

Figure 2: Targeted adversarial examples generated by GSE. The target is airship for the first two rows, and golf cart for the last two rows. The attacked model is a VGG19.

Interpretability Metrics

- $Z(\mathbf{x})$ logits of vectorized image $\mathbf{x} \in [I_{\text{min}}, I_{\text{max}}]^d$
- *Adversarial Saliency Map (ASM)*

$$\text{ASM}(\mathbf{x},t)[i] = \left(\frac{\partial Z(\mathbf{x})_t}{\partial \mathbf{x}_i}\right)\left|\frac{\partial Z(\mathbf{x})_l}{\partial \mathbf{x}_i}\right| 1_S(i),$$

$$S = \left\{ i \in \{1, ..., d\} \,\middle|\, \frac{\partial Z(\mathbf{x})_t}{\partial \mathbf{x}_i} \geq 0 \text{ or } \frac{\partial Z(\mathbf{x})_l}{\partial \mathbf{x}_i} \leq 0 \right\}.$$

- Binary mask $\mathbf{B}_{ASM} \in \{0,1\}^d$

$$\mathbf{B}_{ASM}[i] = \begin{cases} 1, & \text{if ASM}(\mathbf{x},t)[i] > \nu, \\ 0, & \text{otherwise}, \end{cases}$$

- *Interpretability score (IS)* given perturbation $\delta \in \mathbb{R}^d$

$$\text{IS}(\delta) = \frac{\|\mathbf{B}_{ASM} \odot \delta\|_2}{\|\delta\|_2}.$$

- $f_k[i,j]$ activation of the unit k at the coordinates (i,j) in the last convolutional layer
- w_k^l weights corresponding to label l for unit k
- *Class activation map* $\text{CAM}_l[i,j] = \sum_k w_k^l f_k[i,j]$

Quantitative Interpretability

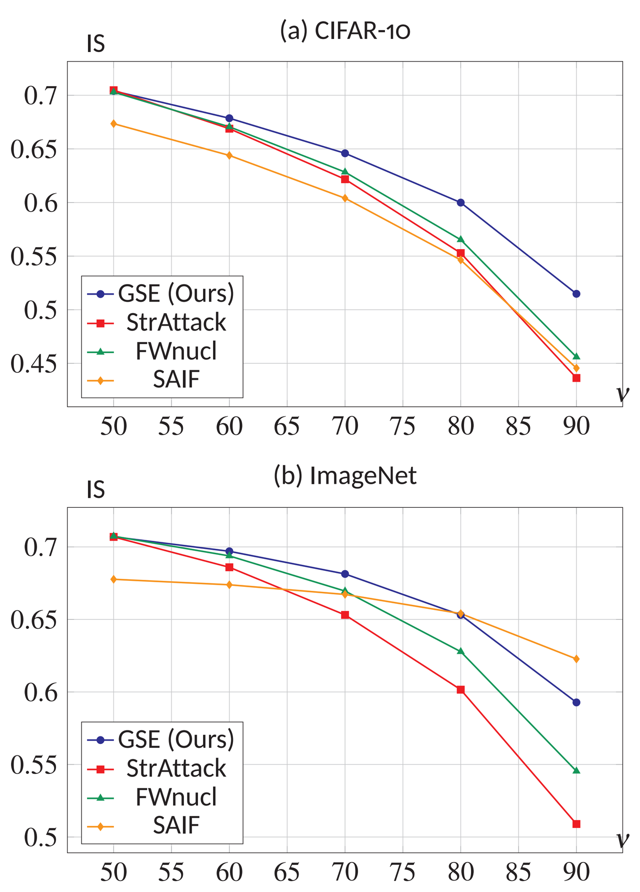

Figure 3: IS vs. percentile ν for targeted GSE attack (1), StrAttack (2), FWnucl (3), and SAIF (4). Evaluated on a CIFAR-10 ResNet20 classifier (a), and an ImageNet VGG19 classifier (b).

Literature

Bibliography

[1] S. Sadiku, M. Wagner, and S. Pokutta. Group-wise Sparse and Explainable Adversarial Attacks. *arXiv preprint arXiv:2311.17434*, 2023.

[2] K. Xu, S. Liu, P. Zhao, P. Chen, H. Zhang, Q. Fan, D. Erdogmus, Y. Wang and X. Lin. StrAttack: Towards general implementation and better interpretability. *ICLR*, 2019.

[3] E. Kazemi and T. Kerdreux and L. Wang. Minimally Distorted Structured Adversarial Attacks *International Journal of Computer Vision 131.1*, pp. 160-176, 2023.

[4] T. Imtiaz, M. Kohler, J. Miller, Z. Wang, M. Sznaier, O. Camps and J. Dy. SAIF: Sparse Adversarial and Interpretable Attack Framework. *arXiv preprint arXiv:2212.07495*, 2022.