# Group-wise Sparse Adversarial Attacks

S. Sadiku, M. Wagner, S. Pokutta

ZUSE INSTITUTE BERLIN

Berlin Mathematical School BMS

## Motivation

- *Deep Neural Networks (DNN)* are vulnerable to adversarial attacks
- *Sparse adversarial attacks* explore $\ell_p$ neighborhoods with $p = 0$ via
  1. Greedy single-pixel selection
  2. Local search techniques
  3. Evolutionary algorithms
  4. Relaxing $\ell_0$ via the $\ell_1$ ball
- Sparse 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

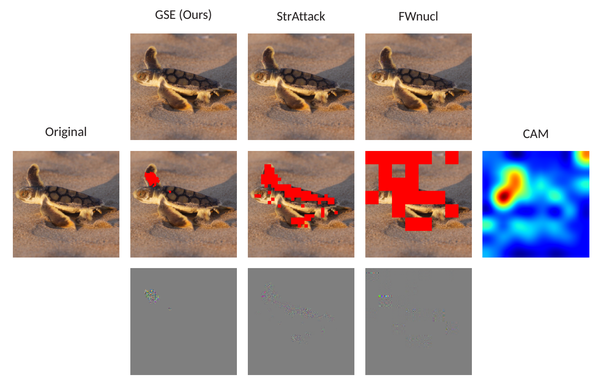


Figure 1: Visual comparison of successful untargeted adversarial instances generated by our attack (1), StrAttack (2), and FWnucl (3). The target model is a ResNet50.

## GSE Adversarial Attacks

- $\mathscr{X} = [I_{\min}, I_{\max}]^{M\times N\times C}$ 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

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

- Closed-form solution for $p = 1/2$ (and $p = \{0, 2/3\}$)
- Tune $\lambda$ 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

## Evaluation metrics and Results on Untargeted Attacks

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

$$\text{ACP} = \frac{1}{mMN}\sum_{i=1}^{m} \|\mathbf{m}^{(i)}\|_0,$$

- Run depth-first search (DFS) on $\mathbf{m}$ 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 $\mathbf{w}$

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

| | Attack | ASR | ACP | ANC | $\ell_2$ | $d_{2,0}$ |
|---|---|---|---|---|---|---|
| CIFAR-10 ResNet20 | GSE (Ours) | **100%** | **41.7** | **1.66** | **0.80** | **177** |
| | StrAttack | **100%** | 118 | 7.50 | 1.02 | 428 |
| | FWnucl | 94.6% | 460 | 1.99 | 2.01 | 594 |
| ImageNet ResNet50 | GSE (Ours) | **100%** | **1629** | 8.42 | **1.50** | **3428** |
| | StrAttack | **100%** | 7265 | 15.3 | 2.31 | 11693 |
| | FWnucl | 47.4% | 13760 | **3.79** | 1.81 | 16345 |
| ImageNet ViT_B_16 | GSE (Ours) | **100%** | **941** | **5.11** | **1.95** | **1964** |
| | StrAttack | **100%** | 3589 | 10.8 | 2.03 | 8152 |
| | FWnucl | 57.9% | 7515 | 5.67 | 3.04 | 9152 |

- ResNet20 classifier for CIFAR-10
- ResNet50 and ViT_B_16 classifiers for ImageNet
- Tested on 1,000 samples from each dataset

## Results on Targeted Attacks

- ResNet20 classifier for CIFAR-10
- ResNet50 and ViT_B_16 classifiers for ImageNet
- Tested on 1,000 samples from each dataset

| | | Best case | | | | | Average case | | | | | Worst case | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Attack | ASR | ACP | ANC | $\ell_2$ | $d_{2,0}$ | ASR | ACP | ANC | $\ell_2$ | $d_{2,0}$ | ASR | ACP | ANC | $\ell_2$ | $d_{2,0}$ |
| CIFAR-10 ResNet20 | GSE (Ours) | **100%** | **29.6** | **1.06** | **0.68** | **137** | **100%** | **86.3** | **1.76** | **1.13** | **262** | **100%** | **162** | **3.31** | **1.57** | **399** |
| | StrAttack | **100%** | 78.4 | 4.56 | 0.79 | 352 | **100%** | 231 | 10.1 | 1.86 | 534 | **100%** | 406 | 15.9 | 4.72 | 619 |
| | FWnucl | **100%** | 283 | 1.18 | 1.48 | 515 | 85.8% | 373 | 2.52 | 2.54 | 564 | 40.5% | 495 | 4.27 | 3.36 | 609 |
| ImageNet ResNet50 | GSE (Ours) | **100%** | **3516** | 5.89 | 2.16 | **5967** | **100%** | **12014** | 14.6 | **2.93** | **16724** | **100%** | **21675** | **22.8** | **3.51** | **29538** |
| | StrAttack | **100%** | 6579 | 7.18 | 2.45 | 9620 | **100%** | 15071 | 18.0 | 3.97 | 20921 | **100%** | 26908 | 32.1 | 6.13 | 34768 |
| | FWnucl | 31.1% | 9897 | **3.81** | **2.02** | 11295 | 7.34% | 19356 | **7.58** | 3.17 | 26591 | 0.0% | N/A | N/A | N/A | N/A |
| ImageNet ViT_B_16 | GSE (Ours) | **100%** | **916** | **3.35** | 2.20 | **1782** | **100%** | **2667** | **7.72** | **2.87** | **4571** | **100%** | **5920** | **14.3** | **3.60** | **9228** |
| | StrAttack | **100%** | 3550 | 7.85 | **2.14** | 5964 | **100%** | 8729 | 17.2 | 3.50 | 13349 | **100%** | 16047 | 27.4 | 5.68 | 22447 |
| | FWnucl | 53.2% | 5483 | 4.13 | 2.77 | 6718 | 11.2% | 6002 | 9.73 | 3.51 | 7427 | 0.0% | N/A | N/A | N/A | N/A |

## Speed Comparison

| | Untargeted | | | Targeted | | |
|---|---|---|---|---|---|---|
| | CIFAR-10 | ImageNet | | CIFAR-10 | ImageNet | |
| Attack | ResNet20 | ResNet50 | ViT_B_16 | ResNet20 | ResNet50 | ViT_B_16 |
| GSE (Ours) | **0.13s** | **3.18s** | **7.05s** | **0.16s** | **4.42s** | **10.3s** |
| StrAttack | 0.97s | 15.2s | 33.6s | 1.28s | 18.5s | 34.2s |
| FWnucl | 0.25s | 9.93s | 26.0s | 0.32s | 11.6s | 26.2s |

## Visual Analysis

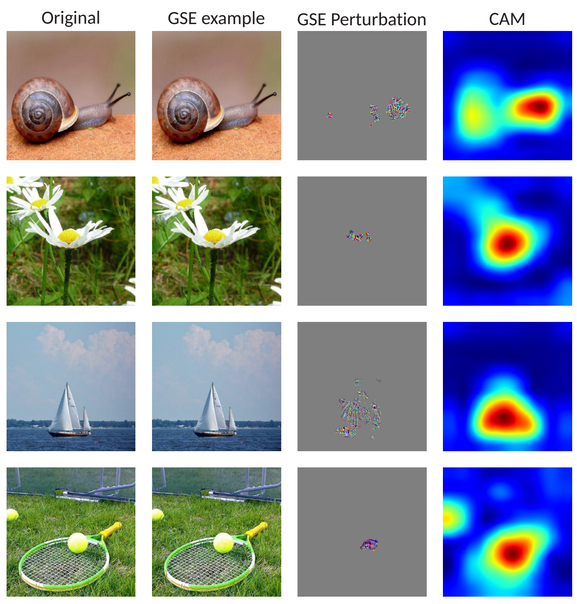


Figure 2: Targeted adversarial examples generated by GSE. The target is airship for the first two rows, and golf cart for the last two rows. The attacked model is a VGG19.

## Interpretability Metrics

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

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

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

- Binary mask $\mathbf{B}(\mathbf{x}, l, t) \in \{0,1\}^d$

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

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

$$\text{IS}(\mathbf{w}, \mathbf{x}, l, t) = \frac{\|\mathbf{B}(\mathbf{x}, l, t) \odot \mathbf{w}\|_2}{\|\mathbf{w}\|_2}$$

- $f_k[i, j]$ 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

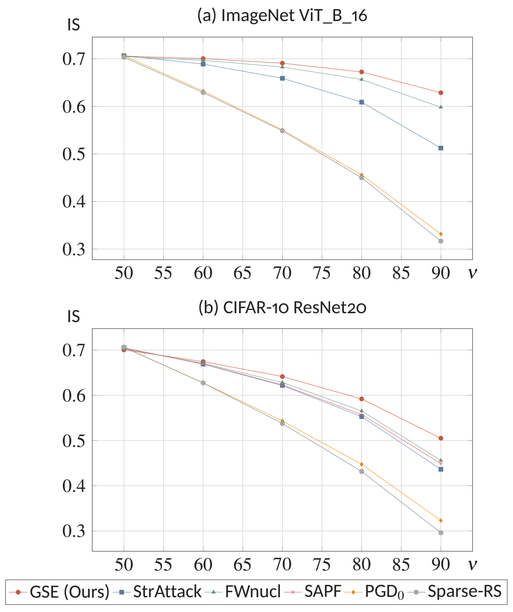


Figure 3: IS vs. percentile $v$ for targeted versions of GSE vs. five other attacks. Evaluated on an ImageNet ViT_B_16 classifier (a), and CIFAR-10 ResNet20 classifier (b).

## Literature

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

[4] Y. Fan, B. Wu, W. Li, Y. Zhang, M. Li, Z. Li and Y. Yang. Sparse adversarial attack via perturbation factorization. *Proceedings of European Conference on Computer Vision*, 2020.