GSE: Group-wise Sparse and Explainable Adversarial Attacks

13th International Conference on Learning Representations (ICLR25)

S. Sadiku, M. Wagner and S. Pokutta

April 2025

Why do we impose structure in adversarial attacks?

- $\mathcal{L} : \mathcal{X} \times \mathbb{N} \to \mathbb{R}$ classification loss function
- Benign image $\boldsymbol{x} \in \mathcal{X}$ of correct label $l \in \mathbb{N}$ and target label $t \in \mathbb{N}, t \neq l$
- Goal of a traditional targeted adversary - succeed under minimal distortion

$$\min_{\boldsymbol{w} \in \mathbb{R}^d} \mathcal{L}(\boldsymbol{x} + \boldsymbol{w}, t) + \lambda \|\boldsymbol{w}\|_p^p$$

  for $\lambda > 0$ and $p \geq 0$

  1. $0 \leq p \leq 1$ changes very few pixels at high magnitudes
     $\hookrightarrow$ Easily **perceptible** even for the human eye (Fan et al., 2020)
  2. $p > 1$ changes most of the pixels at low magnitudes
     $\hookrightarrow$ Appear as **noise** to humans but as features to DNNs (Ilyas et al., 2020))
- Our goal - bridge the gap between human perception and machine interpretation by generating attacks that are
  - **Imperceptible** - low magnitude
  - Targeted at the most **important regions** of the image

Choose key group-wise sparse pixel coordinates

- Consider a vector of tradeoff parameters $\lambda \in \mathbb{R}_{\geq 0}^{M \times N \times C}$
- Heuristically select group-wise sparse coordinates to perturb
  1. Build a mask $\boldsymbol{m} = \text{sign}\left(\sum_{c=1}^{C} |\boldsymbol{w}^{(k)}|_{:,:,c}\right) \in \{0,1\}^{M \times N}$
  2. Apply Gaussian Blur Kernel $\boldsymbol{M} = \boldsymbol{m} ** \boldsymbol{K} \in [0,1]^{M \times N}$
  3. Construct $\overline{\boldsymbol{M}} \in \mathbb{R}^{M \times N}$ via

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

     for $0 < q \leq 1$
  4. Set

$$\lambda_{i,j,:}^{(k+1)} = \frac{\lambda_{i,j,:}^{(k)}}{\overline{\boldsymbol{M}}_{i,j}}$$

$\hookrightarrow$ Denote the chosen pixel coordinates by V

Solve a low magnitude adversarial attack over V

- Formulate a simplified optimization problem

$$\min_{\boldsymbol{w} \in V} \mathcal{L}(\boldsymbol{x} + \boldsymbol{w}, t) + \mu \|\boldsymbol{w}\|_2 \tag{1}$$

  — $\mu > 0$ controls perturbation magnitude
  — Use projected Nesterov's accelerated gradient descent (NAG) to solve Eq. (1)

Proposition (S., Wagner and Pokutta, 2025)

The projected NAG solving Eq. (1) converges as NAG solving an unconstrained problem.

- Contrary to benchmarks, our method does not depend on pixel partitionings

Results on targeted adversarial attacks

Table: Targeted attacks performed on ResNet20 classifier for CIFAR-10, and ResNet50 and ViT_B_16 classifiers for ImageNet. Tested on 1k images from each dataset, 9 target labels for CIFAR-10 and 10 target labels for ImageNet.

		Best case					Average case					Worst case				
	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**	**1.06**	**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

Quantitative evaluation

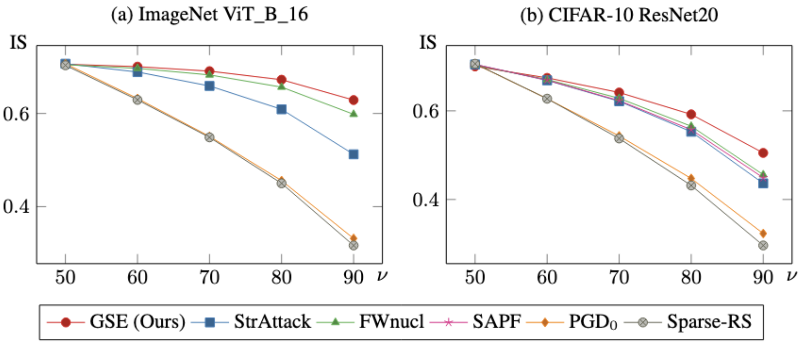

Figure: IS vs. percentile ν for targeted versions of GSE vs. five other attacks. Evaluated on an ImageNet ViT_B_16 classifier (a), and CIFAR-10 ResNet20 classifier (b). Tested on 1k images from each dataset, 9 target labels for CIFAR-10 and 10 target labels for ImageNet.

Thank you for your attention!