GSE: Group-wise Sparse and Explainable Adversarial Attacks

Shpresim Sadiku, Moritz Wagner, Sebastian Pokutta

Cooperation: TU Berlin, ZIB

Funding: DFG Cluster of Excellence Math+, German Federal Ministry of Education and Research

Berlin Mathematics Research Center MATH+

ZIB ZUSE INSTITUTE BERLIN

Motivation

- *Sparse adversarial attacks* often produce perturbations that are ambiguous about which regions of the image are important for classification
- Group-wise sparse methods often lead to reduced ambiguity about the salient regions in an image but
  - rely on predefined pixel partitionings
  - produce less sparse perturbations
- Generate imperceptible, group-wise sparse adversarial attacks that target the image's main objective, ensuring *explainable* perturbations without pixel partitioning or loss of sparsity

GSE Adversarial Attacks

$\mathscr{X} = [I_{\min}, I_{\max}]^{M\times N\times C}$ is the set of feasible images and $\mathscr{L} : \mathscr{X} \times \mathbb{N} \to \mathbb{R}$ a classification loss function

- *Targeted sparse adversarial attacks* find a perturbation w for given image x and target t via

$$\min_{w\in\mathbb{R}^{M\times N\times C}} \mathscr{L}(x+w,t) + \lambda\|w\|_p^p \qquad (1)$$

- Solve (1) using forward-backward splitting for $p \in (0,1)$ with per-pixel trade-off parameter λ
- For $p = \frac{1}{2}$, there exists a closed-form solution for the proximal operator

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

- Heuristically impose a group-sparsity structure by tuning each pixel's λ depending on its proximity to an already perturbed pixel via blurring

$$\lambda_{i,j}^{(k+1)} = \frac{\lambda_{i,j}^{(k)}}{\overline{M}_{i,j}}, \quad \overline{M}_{i,j} = \begin{cases} M_{i,j}+1 & \text{if } M_{i,j} \neq 0 \\ q, & \text{else} \end{cases}$$

$$M_{i,j} = \text{sign}\left(\sum_{c=1}^{C} |w^{(k)}|_{:,:,c}\right) ** K$$

- After $\tilde{k}$ iterations, solve (1) with $p = 2$, constrained to the set of pixels (i,j) with $\lambda_{i,j}^{(\tilde{k})} < \lambda_{i,j}^{(1)}$ using Nesterov's Accelerated Gradient Method

Evaluation metrics and Results on Untargeted Attacks

- $(x^{(i)})_{0<i\leq n}$ images of perturbation $(w^{(i)})_{0<i\leq n}$
- *Attack Success Rate* ASR $= \frac{m_s}{n}$ for m_s successful adversaries
- *Average Number of Changed Pixels (ACP)* - $\frac{1}{m_s MN}\sum_{i=1}^{m_s}\|m^{(i)}\|_0$
- *Average Number of Clusters (ANC)* - the number of connected clusters of perturbed pixels averaged over all successful attacks
- Group-wise sparsity measure for a set $\{G_1,\ldots,G_k\}$ of overlapping n-by-n pixel patches $d_{2,0}(w) := |\{i \,:\, \|w_{G_i}\|_2 \neq 0,\ i = 1,\ldots,k\}|$

	Attack	ASR	ACP	ANC	ℓ_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

Results on Targeted Attacks

		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

Visual Analysis

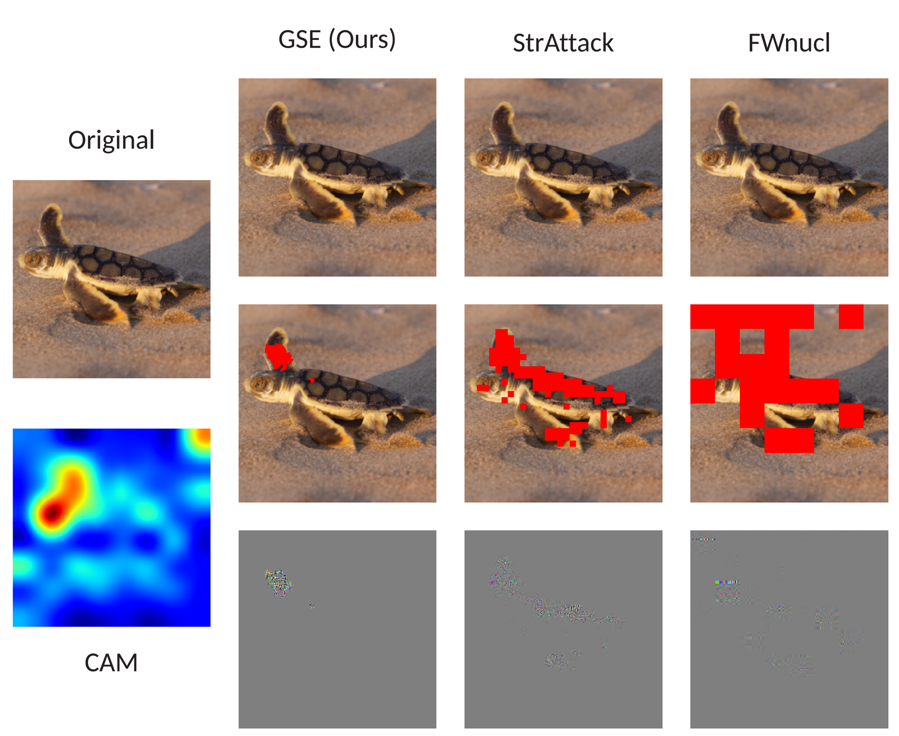

Figure 1: Visual comparison of successful untargeted adversarial instances generated by our attack, StrAttack, and FWnucl. The attacked model is a ResNet50.

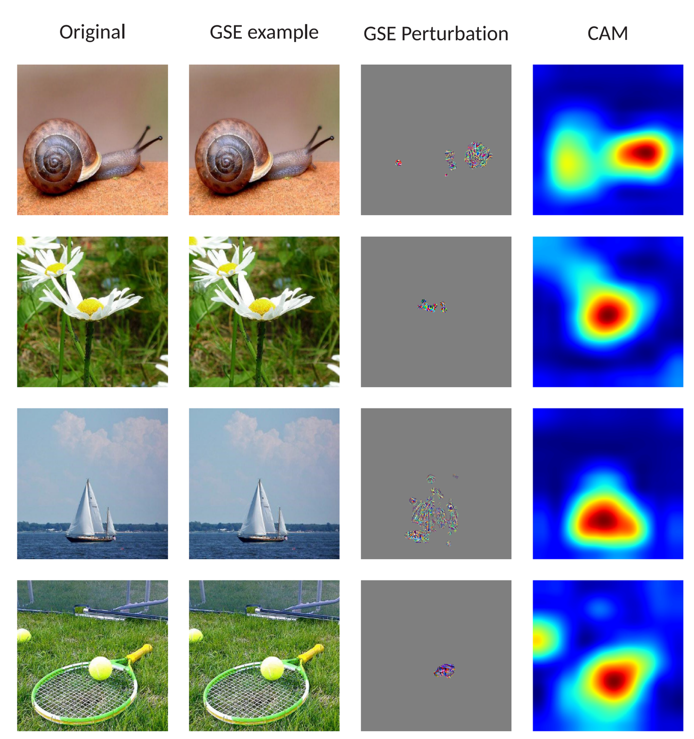

Figure 2: Targeted adversarial examples generated by GSE. The target is airship for the first two rows, and golf cart for the last two rows. The attacked model is a VGG19.

Interpretability Metrics

$Z(x)$ are the logits of the vectorized image $x \in [I_{\min}, I_{\max}]^d$, l is the true label, and t a target label

- Use the *Interpretability score* (IS) for quantitative analysis. For a given perturbation $w \in \mathbb{R}^d$

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

based on the *Adversarial Saliency Map (ASM)* [1], where

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

- Utilize *Class activation map (CAM)* [2] for qualitative interpretability analysis

Interpretability Quantitatively

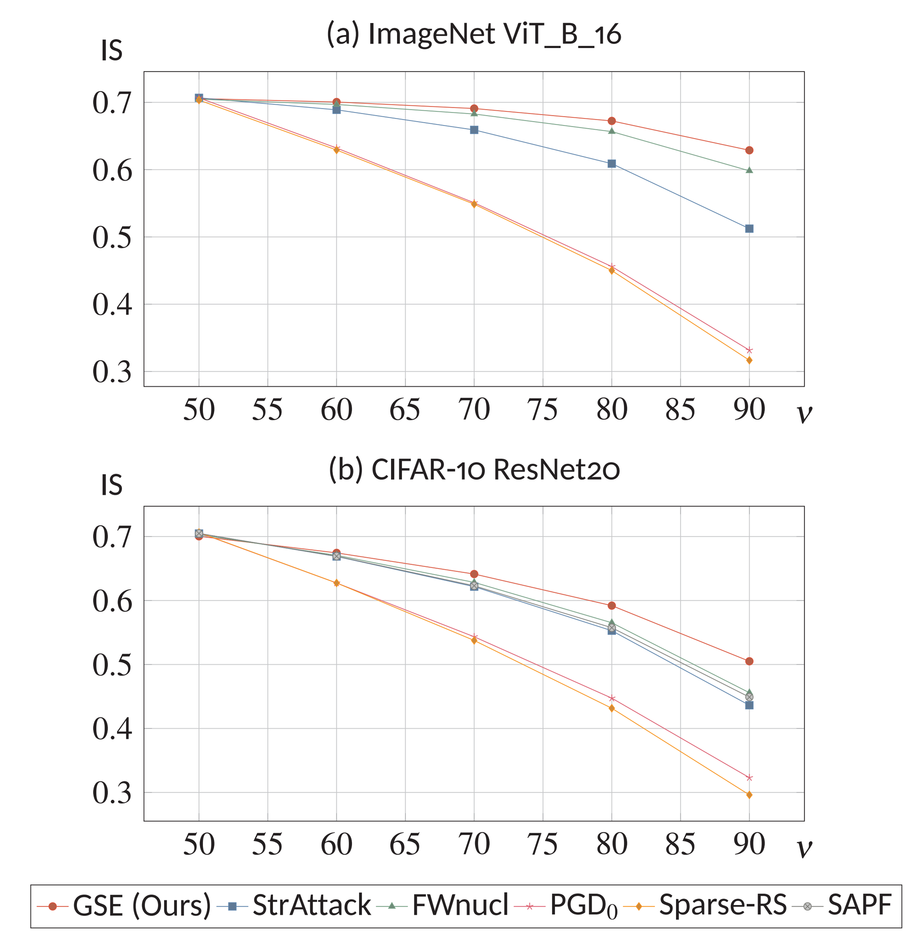

Figure 3: IS vs. percentile ν for targeted versions of GSE vs. five other attacks. Evaluated on an ImageNet ViT_B_16 classifier (a), and CIFAR-10 ResNet20 classifier (b).

References

[1] N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, and A. Swami. The limitations of deep learning in adversarial settings. *IEEE European symposium on security and privacy*, 2016.

[2] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. Learning deep features for discriminative localization. *IEEE conference on computer vision and pattern recognition*, 2016.