Detecting Adversarial Data Using Perturbation Forgery

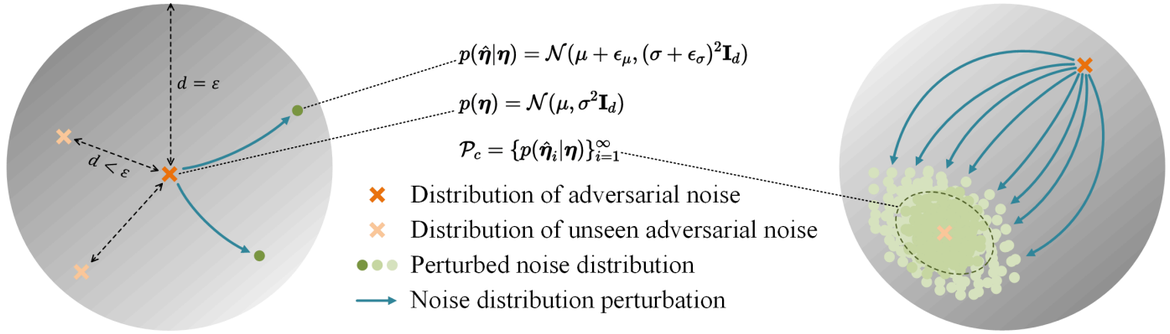
Qian Wang, Chen Li, Yuchen Luo, Hefei Ling, Shijuan Huang, Ruoxi Jia, Ning Yu

Introduction

- We explore the proximity among adversarial noise distributions and demonstrate the existence of an open covering.
- Training on this open covering enables the development of a detector that generalizes well to unseen attacks.
- The detector introduces minimal inference-time overhead.

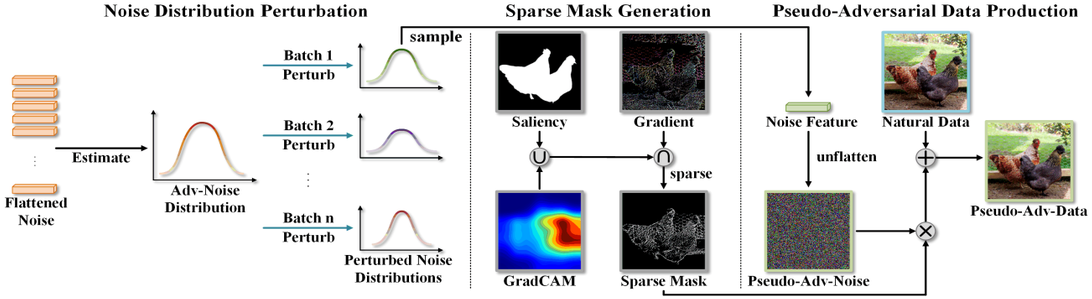
Methods	Detect Attacks	Model-Agnostic	Unseen-Attack Detection AUC	Time Overhead
LID	Gradient	×	0.9146	1.80s
LiBRe	Gradient	×	0.8725	2.56s
SPAD	Gradient + GAN	✓	0.9820	4.56s
EPSAD	Gradient	✓	0.9918	396.81s
Ours	Gradient + GAN + Diffusion	✓	0.9931	4.85

Proximity of Adversarial Noise Distribution



Left: All adversarial noise distributions and their perturbations reside within an ε-ball centered at a given adversarial noise distribution.
Right: An open covering of adversarial noise distributions is obtained through continuous perturbations of the given distribution.

Perturbation Forgery



1. Estimate the noise distribution induced by a commonly used attack before training.
2. Continuously perturb it across batches to form an open covering.
3. Convert sampled noise into localized noise by applying sparse masks.
4. Generated pseudo-adversarial data by adding localized noise to natural samples.

Detection

Detecting Gradient-based Adversarial Attack:

Detector	BIM	PGD	DIM	MIM	NIM	AA	BPDA+EOT	MM
LID	0.9782	0.9750	0.8942	0.9146	0.8977	0.9124	0.9202	0.9227
LiBRe	0.9259	0.9548	0.9243	0.8725	0.9013	0.8653	0.8714	0.8573
SPAD	0.9846	0.9851	0.9815	0.9820	0.9823	0.9890	0.9885	0.9811
EPSAD	**0.9998**	**0.9989**	**0.9923**	0.9918	**0.9972**	**0.9998**	**0.9985**	0.9923
ours	0.9911	0.9912	0.9863	**0.9931**	0.9934	0.9941	0.9927	**0.9944**

Detecting Generative-based Adversarial Attack:

Detector	CDA	TTP	M3D	Diff-Attack	Diff-PGD
CNN-Detection	0.7051	0.6743	0.6917	0.3963	0.5260
LGrad	0.6144	0.6077	0.6068	0.5586	0.5835
Universal-Detector	0.7945	0.8170	0.8312	0.9202	0.5531
DIRE	0.8976	0.9097	0.9129	0.9097	0.9143
SPAD	0.9385	0.9064	0.8984	0.8862	0.8879
EPSAD	0.9674	0.6997	0.7265	0.4700	0.1507
ours	**0.9878**	**0.9678**	**0.9364**	**0.9223**	**0.9223**