Adversarial Purification with Score-based Generative Models

Jongmin Yoon, Sung Ju Hwang, Juho Lee
KAIST

Adversarial purification

Adversarial attack

- An image containing a *small perturbation to human* completely changes the prediction results

Adversarial training

- Train a neural network with *adversarial images*

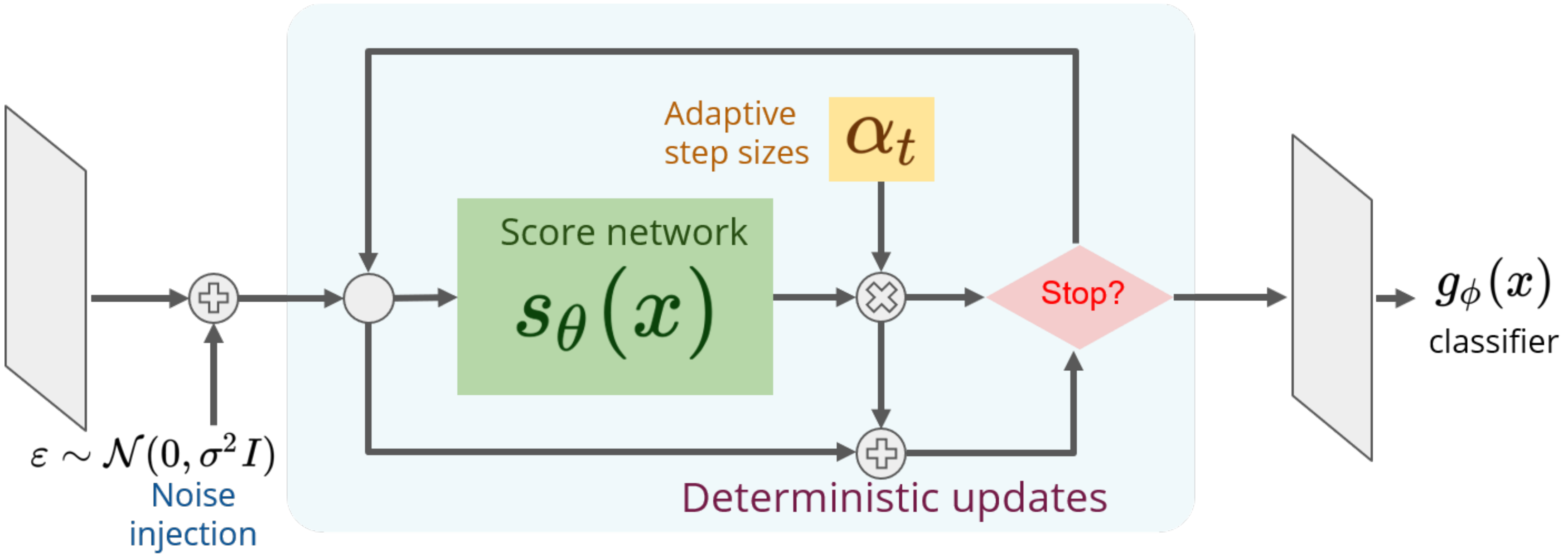
Adversarial purification

- Defend a trained classifiers by using an ***additional purifier network***
- Consider purification as denoising of the adversarial attacks

Adaptive Denoising Purification (ADP)

Our defense strategy, ADP, consists of 3 steps:

- Screening attacked images by random noise
- Purification by deterministic updates
- Merging duplicate of purified images and predict



Adaptive Denoising Purification **(ADP)**

Our defense strategy, ADP, consists of 3 steps:

- Screening attacked images by random noise
- Purification by deterministic updates
- Merging duplicate of purified images and predict

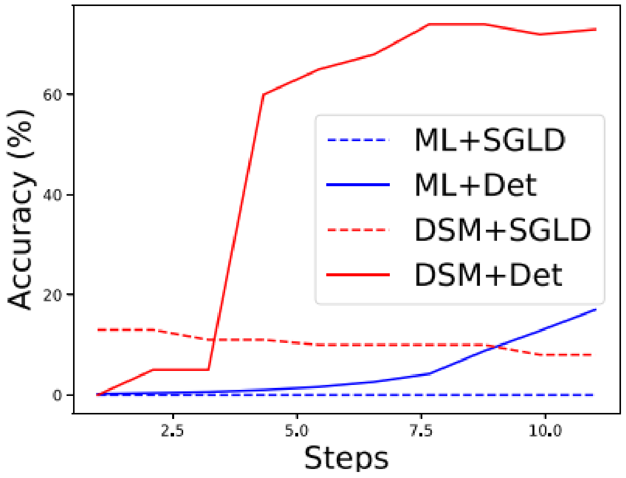
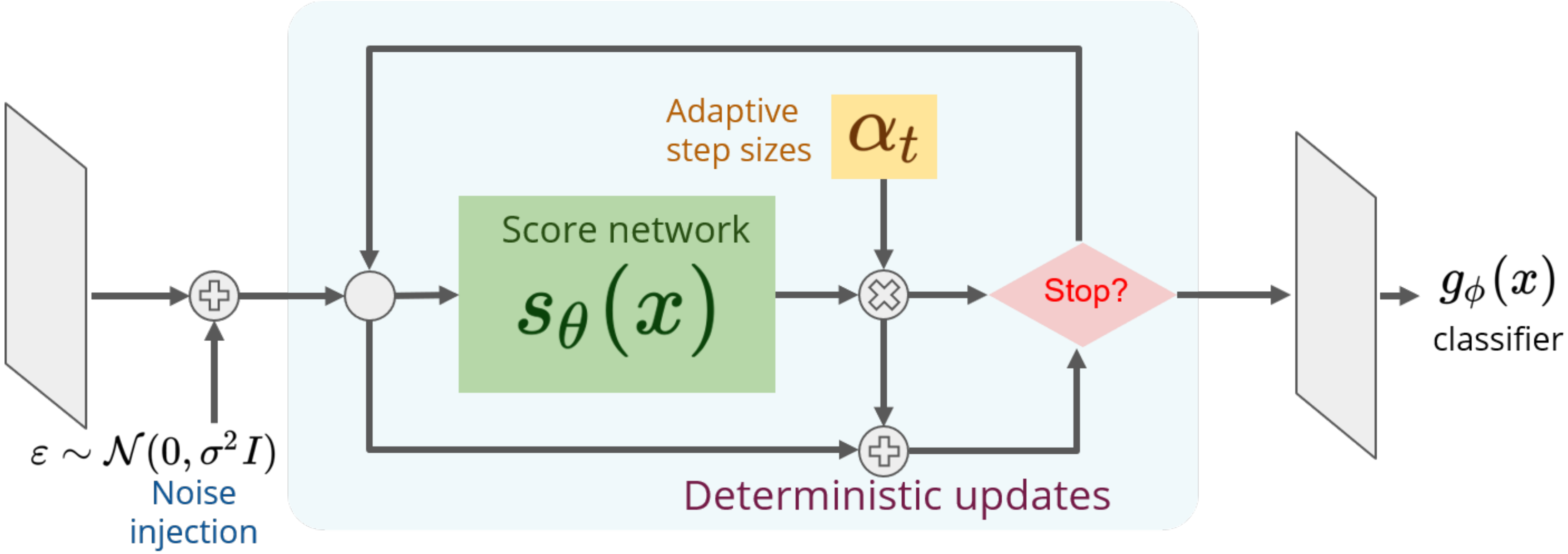


Figure 2. The accuracy against the BPDA attack on CIFAR10. ML denotes the maximum likelihood training with MCMC, and Det denotes deterministic updates.

Adaptive Denoising Purification **(ADP)**

Our defense strategy, ADP, consists of 3 steps:

- Screening attacked images by random noise
- Purification by deterministic updates
- Merging duplicate of purified images and predict



Adversarial attacks against ADP

The list of adversarial attacks designed to break ADP

1. Classifier PGD (Preprocessor–blind) [Madry et al., 2015]
2. BPDA+EOT attack (Strong adaptive) [Athalye et al., 2018]
3. SPSA attack (Score–based black–box) [Uesato et al., 2018]

For all attacks, the threat models are fixed to ℓ_∞ ε–ball with $\varepsilon = 8/255$.

Table 1. List of attacks considered. After each update, the output is projected with $x_{i+1} = \prod_{\mathcal{B}_\infty(x_0,b)} x'_{i+1}$. Here $f_\theta : \mathbb{R}^D \to \mathbb{R}^D$ is the full purification model, $s_\theta : \mathbb{R}^D \to \mathbb{R}^D$ is the score network that consists the purification and $g_\phi : \mathbb{R}^D \to \mathbb{R}^K$ is the classifier, where D is the dimension of data and K is the number of classes. For SPSA attack, v_i is uniformly sampled from $\{-1, 1\}^D$. For all of our experiments, we fix $\alpha_i = 2/255$ and $\varepsilon = 0.5$.

Attack name	Type	Updating rule to derive x'_{i+1}
Full gradient	White-box	$x_i + \alpha_i \mathtt{sign} \nabla_x \mathcal{L}\left(\left(g_\phi \circ f_\theta\right)(x), y\right)\vert_{x=x_i}$
Classifier PGD	Preprocessor-blind	$x_i + \alpha_i \mathtt{sign} \nabla_x \mathcal{L}\left(g_\phi(x), y\right)\vert_{x=x_i}$
BPDA (Athalye et al., 2018)	Adaptive	$x_i + \alpha_i \mathtt{sign} \nabla_x \mathcal{L}\left(g_\phi(x), y\right)\vert_{x=f_\theta(x_i)}$
Joint attack (score)	Adaptive	$x_i + \alpha_i\left(\varepsilon \mathtt{sign}(s_\theta(x_i)) + (1-\varepsilon)\mathtt{sign}(\nabla_x \mathcal{L}(g_\phi(x), y)\vert_{x=x_i}\right)$
Joint attack (full)	Adaptive	$x_i + \alpha_i\left(\varepsilon \mathtt{sign}(f_\theta(x_i) - x_i) + (1-\varepsilon)\mathtt{sign} \nabla_x \mathcal{L}(g_\phi(x), y)\vert_{x=x_i}\right)$
SPSA (Uesato et al., 2018)	Black-box	$x_i + \alpha_i \mathtt{sign} \sum_{j=1}^{N} \frac{\mathcal{L}(((g_\phi \circ f_\theta)(x+\varepsilon v_j), y) - \mathcal{L}((g_\phi \circ f_\theta)(x-\varepsilon v_j), y)) \cdot v_j}{2N\varepsilon}$

Experiment results

CIFAR-10, Strong adaptive attack

Models / Attacks	Accuracy (%) Natural	Robust	Architecture
ADP ($\sigma = 0.25$)	86.14		
BPDA 40+EOT		**70.01**	WRN-28-10
BPDA 100+EOT		**69.71**	WRN-28-10
Joint (score)+EOT		70.61	WRN-28-10
Joint (full)+EOT		78.39	WRN-28-10
SPSA		80.80	WRN-28-10
Adversarial purification methods			
(Hill et al., 2021)	84.12	54.90	WRN-28-10
(Song et al., 2018)*	95.00	9	ResNet-62
(Yang et al., 2019)*	88.7	55.1	WRN-28-10
(Shi et al., 2021)*	91.89	53.58	WRN-28-10
Adversarial training methods			
(Madry et al., 2018)*	87.3	45.8	ResNet-18
(Zhang et al., 2019)*	84.90	56.43	ResNet-18
(Carmon et al., 2019)	89.67	63.1	WRN-28-10
(Gowal et al., 2020)*	89.48	64.08	WRN-28-10

CIFAR-10, Preprocessor-blind attack

Models	Accuracy (%) Standard	Robust	Architecture
Raw WideResNet	95.80	0.00	WRN-28-10
ADP ($\sigma = 0.1$)	93.09	85.45	WRN-28-10
ADP ($\sigma = 0.25$)	86.14	80.24	WRN-28-10
Adversarial purification methods			
(Hill et al., 2021)	84.12	78.91	WRN-28-10
(Shi et al., 2021)*	96.93	63.10	WRN-28-10
(Du & Mordatch, 2019)*	48.7	37.5	WRN-28-10
(Grathwohl et al., 2020)*	75.5	23.8	WRN-28-10
(Yang et al., 2019)*			
$p = 0.8 \to 1.0$	94.9	82.5	ResNet-18
$p = 0.6 \to 0.8$	92.1	80.3	ResNet-18
$p = 0.4 \to 0.6$	89.2	77.4	ResNet-18
(Song et al., 2018)*			
Natural + PixelCNN	82	61	ResNet-62
AT + PixelCNN	90	70	ResNet-62
Adversarial training methods, transfer-based			
(Madry et al., 2018)*	87.3	70.2	ResNet-56
(Zhang et al., 2019)*	84.9	72.2	ResNet-56

Conclusion

- EBM trained with denoising score matching quickly purifies attacked images with deterministic short-run updates.
- Our DSM-based purification shows superior performance compared to existing methods.
- Some further directions
    - Certified robustness: As a generative randomized smoothing classifier, further investigation on denoising-based adversarial purification will shed light on certified robustness that can also be achieved empirically. A brief analysis is introduced in our main paper.
    - Scalability: Recent progress on score-based generative modelling and diffusion model can also facilitate adversarial purification for larger-scale images