Diffusion Models for Adversarial Purification

Weili Nie[1], Brandon Guo[2], Yujia Huang[2], Chaowei Xiao[1,3], Arash Vahdat[1], Anima Anandkumar[2]

[1]NVIDIA, [2]Caltech, [3]ASU

Motivation

Adversarial training or adversarial purification?

- **Adversarial training:** *It trains classifiers on adversarial examples*
  - Defense against seen threats
  - Defense against unseen threats
  - Training complexity
- **Adversarial purification:** *It uses generative models to purify adversarial perturbations*
  - Defense against seen threats
  - Defense against unseen threats
  - Training complexity*

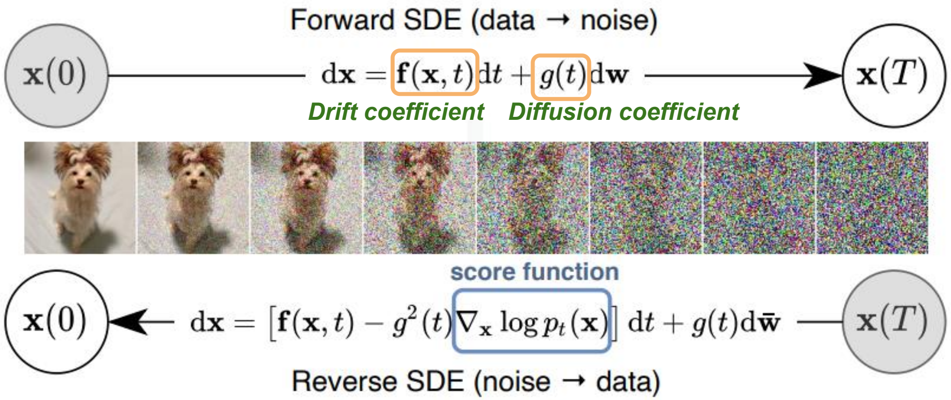
Can we overcome the shortcomings of adversarial purification with <u>a better generative prior</u>?

* It assumes we already have pre-trained generative models.

Motivation

Diffusion models have emerged as powerful generative models

Guided-Diffusion (Dhariwal & Nichol, 2021)

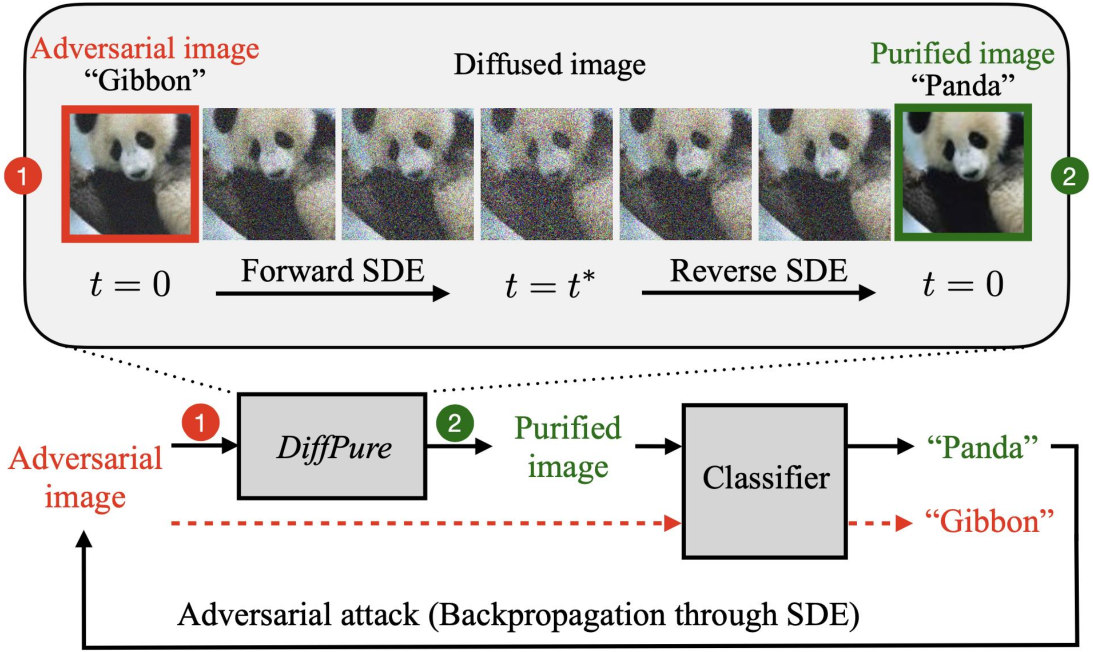


(Song et al., 2021)

How should we use a diffusion model as the purification model for adversarial defense?

DiffPure (Diffusion Purification)

It uses the forward and reverse processes of pre-trained diffusion models to purify adversarial images



How to Evaluate DiffPure on Strong Adaptive Attacks?

We use adjoint method to compute full gradients of reverse SDE for adaptive attacks

- Challenge
  - Strong adaptive attacks (e.g. AutoAttack) require computing full gradients of DiffPure
  - Naively backpropagating through SDE scales poorly in memory
- Our solution
  - Use *adjoint method* to compute gradient of SDE
  - Convert gradient computation to solving an augmented SDE in *Eq. (6)*

Proposition 3.3. *For the SDE in Eq. (4), the augmented SDE that computes the gradient* $\frac{\partial \mathcal{L}}{\partial \boldsymbol{x}(t^*)}$ *of backpropagating through it is given by*

$$\begin{pmatrix} \boldsymbol{x}(t^*) \\ \frac{\partial \mathcal{L}}{\partial \boldsymbol{x}(t^*)} \end{pmatrix} = \texttt{sdeint}\left(\begin{pmatrix} \hat{\boldsymbol{x}}(0) \\ \frac{\partial \mathcal{L}}{\partial \hat{\boldsymbol{x}}(0)} \end{pmatrix}, \tilde{\boldsymbol{f}}, \tilde{\boldsymbol{g}}, \tilde{\boldsymbol{w}}, 0, t^* \right) \tag{6}$$

where $\frac{\partial \mathcal{L}}{\partial \hat{\boldsymbol{x}}(0)}$ *is the gradient of the objective* $\mathcal{L}$ *w.r.t. the output* $\hat{\boldsymbol{x}}(0)$ *of the SDE in Eq. (4), and*

$$\tilde{\boldsymbol{f}}([\boldsymbol{x}; \boldsymbol{z}], t) = \begin{pmatrix} \boldsymbol{f}_{rev}(\boldsymbol{x}, t) \\ \frac{\partial \boldsymbol{f}_{rev}(\boldsymbol{x}, t)}{\partial \boldsymbol{x}} \boldsymbol{z} \end{pmatrix}$$

$$\tilde{\boldsymbol{g}}(t) = \begin{pmatrix} -g_{rev}(t) \mathbf{1}_d \\ \mathbf{0}_d \end{pmatrix}$$

$$\tilde{\boldsymbol{w}}(t) = \begin{pmatrix} -\boldsymbol{w}(1-t) \\ -\boldsymbol{w}(1-t) \end{pmatrix}$$

with $\mathbf{1}_d$ *and* $\mathbf{0}_d$ *representing the d-dimensional vectors of all ones and all zeros, respectively.*

Implemented in the "TorchSDE" library (Li et al., 2020)

Comparison with SOTA in RobustBench Benchmark: CIFAR-10

DiffPure has absolute improvements of up to +5% in robust accuracy

Method	Extra Data	Standard Acc	Robust Acc
WideResNet-28-10			
(Zhang et al., 2020)	✓	89.36	59.96
(Wu et al., 2020)	✓	88.25	62.11
(Gowal et al., 2020)	✓	89.48	62.70
(Wu et al., 2020)	✗	85.36	59.18
(Rebuffi et al., 2021)	✗	87.33	61.72
(Gowal et al., 2021)	✗	87.50	65.24
Ours	✗	**89.02±0.21**	**70.64±0.39**
WideResNet-70-16			
(Gowal et al., 2020)	✓	91.10	66.02
(Rebuffi et al., 2021)	✓	92.23	68.56
(Gowal et al., 2020)	✗	85.29	59.57
(Rebuffi et al., 2021)	✗	88.54	64.46
(Gowal et al., 2021)	✗	88.74	66.60
Ours	✗	**90.07±0.97**	**71.29±0.55**

AutoAttack Linf (eps=8/255)

Method	Extra Data	Standard Acc	Robust Acc
WideResNet-28-10			
(Augustin et al., 2020)*	✓	92.23	77.93
(Rony et al., 2019)	✗	89.05	66.41
(Ding et al., 2020)	✗	88.02	67.77
(Wu et al., 2020)*	✗	88.51	72.85
(Sehwag et al., 2021)*	✗	90.31	75.39
(Rebuffi et al., 2021)	✗	**91.79**	78.32
Ours	✗	91.03±0.35	**78.58±0.40**
WideResNet-70-16			
(Gowal et al., 2020)	✓	94.74	79.88
(Rebuffi et al., 2021)	✓	95.74	81.44
(Gowal et al., 2020)	✗	90.90	74.03
(Rebuffi et al., 2021)	✗	92.41	**80.86**
Ours	✗	**92.68±0.56**	80.60±0.57

AutoAttack L2 (eps=0.5)

Comparison with SOTA in RobustBench Benchmark: ImageNet

DiffPure has absolute improvements of up to +7% in robust accuracy

Method	Extra Data	Standard Acc	Robust Acc
ResNet-50			
(Engstrom et al., 2019)	✗	62.56	31.06
(Wong et al., 2020)	✗	55.62	26.95
(Salman et al., 2020)	✗	64.02	37.89
(Bai et al., 2021)†	✗	67.38	35.51
Ours	✗	**67.79±0.43**	**40.93±1.96**
WideResNet-50-2			
(Salman et al., 2020)	✗	68.46	39.25
Ours	✗	**71.16±0.75**	**44.39±0.95**
DeiT-S			
(Bai et al., 2021)†	✗	66.50	35.50
Ours	✗	**73.63±0.62**	**43.18±1.27**

AutoAttack Linf (eps=4/255)

Defense Against Unseen Threats: CIFAR-10

DiffPure has absolute improvements of up to +36% in robust accuracy

Method	Standard Acc	Robust Acc		
		ℓ_∞	ℓ_2	StAdv
Adv. Training with ℓ_∞ (Laidlaw et al., 2021)	86.8	49.0	19.2	4.8
Adv. Training with ℓ_2 (Laidlaw et al., 2021)	85.0	39.5	47.8	7.8
Adv. Training with StAdv (Laidlaw et al., 2021)	86.2	0.1	0.2	53.9
PAT-self (Laidlaw et al., 2021)	82.4	30.2	34.9	46.4
ADV. CRAIG (Dolatabadi et al., 2021)	83.2	40.0	33.9	49.6
ADV. GRADMATCH (Dolatabadi et al., 2021)	83.1	39.2	34.1	48.9
Ours	**88.2±0.8**	**70.0±1.2**	**70.9±0.6**	**55.0±0.7**

AutoAttack Linf (eps=8/255), **AutoAttack L2** (eps=0.5) and **StAdv** (eps=0.05)

Comparison with Other Purification Methods

DiffPure has absolute improvements of +15% on CelebA-HQ and +11% on CIFAR-10 in robust accuracy

(a) CelebA-HQ

Method	Purification	Standard Acc	Robust Acc
(Vahdat & Kautz, 2020)	VAE	**99.43**	0.00
(Karras et al., 2020)	GAN+OPT	97.76	10.80
(Chai et al., 2021)	GAN+ENC+OPT	99.37	26.37
(Richardson et al., 2021)	GAN+ENC	93.95	75.00
Ours ($t^* = 0.4$)	Diffusion	93.87±0.18	89.47±1.18
Ours ($t^* = 0.5$)	Diffusion	93.77±0.30	**90.63±1.10**

(b) CIFAR-10

Method	Purification	Standard Acc	Robust Acc
(Song et al., 2018)	Gibbs Update	**95.00**	9.00
(Yang et al., 2019)	Mask+Recon.	94.00	15.00
(Hill et al., 2021)	EBM+LD	84.12	54.90
(Yoon et al., 2021)	DSM+LD*	86.14	70.01
Ours ($t^* = 0.075$)	Diffusion	91.03±0.35	77.43±0.19
Ours ($t^* = 0.1$)	Diffusion	89.02±0.21	**81.40±0.16**

BPDA+EOT Linf (eps=16/255 for CelebA-HQ, eps=8/255 for CIFAR-10)

Qualitative Results of DiffPure on CelebA-HQ

DiffPure removes adversarial perturbations on different attribute classifiers