AdvDiff: Generating Unrestricted Adversarial Examples using Diffusion Models

DAI Xuelong, LIANG Kaisheng, XIAO Bin

Presenter: DAI Xuelong

August 21, 2024

Table of Contents

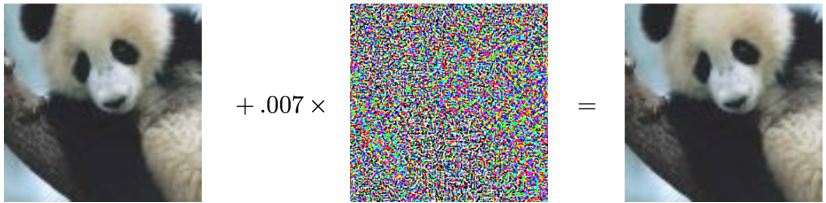
Overview

- AdvDiff is an unrestricted adversarial attack with the reverse generation process of diffusion models.
- AdvDiff utilizes interpretable adversarial guidances to achieve high-quality sampling.
- AdvDiff outperforms SOTA attacks on both image quality and ASR.

Table of Contents

Motivation

- Deep Neural Networks (DNNs) are known to be vulnerable to adversarial attacks.



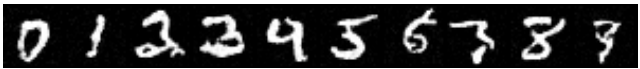
Motivation

- Deep Neural Networks (DNNs) are known to be vulnerable to adversarial attacks.
- However, even with a small perturbation budget, these attacks can be identified by humans.

PGD

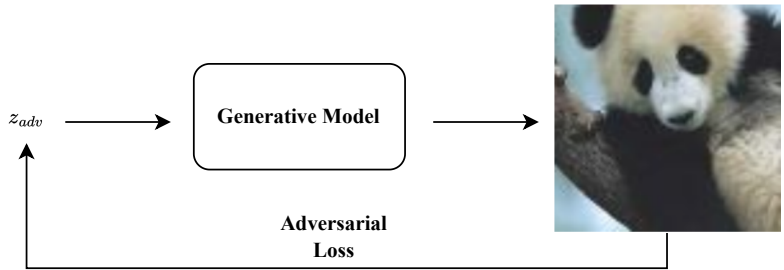
U-GAN

AdvDiff



Motivation

- Unrestricted adversarial attacks aim to generate natural adversarial examples using generative models.



Motivation

- Unrestricted adversarial attacks aim to generate natural adversarial examples using generative models.
- Previous attacks directly inject PGD-like gradient into the sampling of generative models, which harms the generation quality.

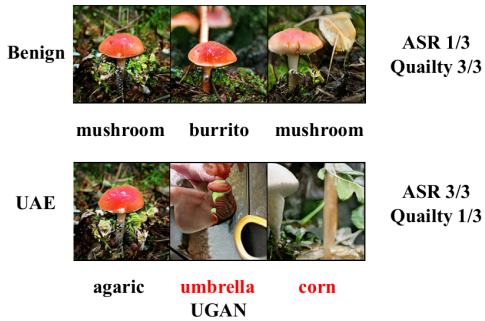


Table of Contents

Contribution

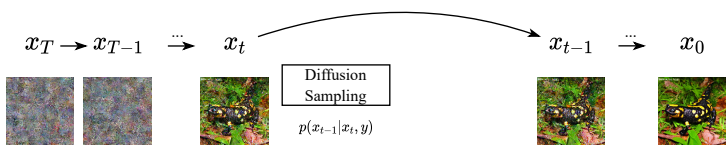
- Two new effective adversarial guidance techniques to the diffusion sampling process that incorporate adversarial objectives to the diffusion model without re-training the model.
- Theoretical analysis reveals that AdvDiff can generate unrestricted adversarial examples while preserving the high-quality and stable sampling of the conditional diffusion models.
- Supports both DDIM and DDPM sampler and is easy to transfer to any diffusion model.

Table of Contents

Method: Adversarial Guidance

- Benign diffusion models reverse generation process recovers the data x_0 with a sequence of noisy data $\{x_{T-1}, \ldots, x_1\}$ by sampling from $p_\theta(x_{t-1}|x_t)$.
- Our aim is to generate the adversarial examples with the target label y_a for conditional sampling.

$$p(x^*_{t-1}|x_t, y_a) = \frac{p(y_a|x^*_{t-1}, x_t)p(x^*_{t-1}|x_t)}{p(y_a|x_t)} \tag{1}$$



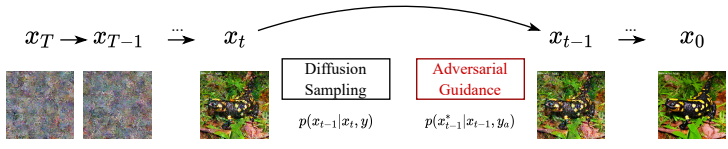
Reverse Generation Process with AdvDiff

Method: Adversarial Guidance

- Assume $p(x^*_{t-1}|x_t) = \mathcal{N}(x^*_{t-1}; \mu(x_t), \sigma_t^2\mathbf{I}) \propto e^{-(x^*_{t-1}-\mu(x_t))^2/2\sigma_t^2}$, we can deduce:

$$x^*_{t-1} = \mu(x_t, y) + \sigma_t\varepsilon + \sigma_t^2 s\nabla_{\mu(x_t)} \log p_f(y_a|\mu(x_t)) \approx x_{t-1} + \sigma_t^2 s\nabla_{x_{t-1}} \log p_f(y_a|x_{t-1}) \quad (2)$$

- The first term represents benign diffusion sampling, and the second term represents adversarial guidance.



Reverse Generation Process with AdvDiff

Method: Noise Sampling Guidance

- We further improve the reverse process by adding an adversarial label prior to the noise data x_T.

$$
\begin{aligned}
p(x_T|y_a) =& \frac{p(y_a|x_T)p(x_T)}{p(y_a)} = \frac{p(y_a|x_T, x_0)p(x_T|x_0)}{p(y_a|x_0)} \\
=& p(x_T|x_0)e^{\log p(y_a|x_T) - \log p(y_a|x_0)}
\end{aligned} \tag{3}
$$

- We can infer the x_T with the adversarial prior, i.e.,

$$
x_T = (\mu(x_0, y) + \sigma_t \varepsilon) + \bar{\sigma}_T^2 a \nabla_{x_0} \log p_f(y_a|x_0) \tag{4}
$$

Method: Noise Sampling Guidance

- We can infer the x_t with the adversarial prior, i.e.,

$$x_T = (\mu(x_0, y) + \sigma_t \varepsilon) + \bar{\sigma}_T^2 a \nabla_{x_0} \log p_f(y_a | x_0) \tag{5}$$

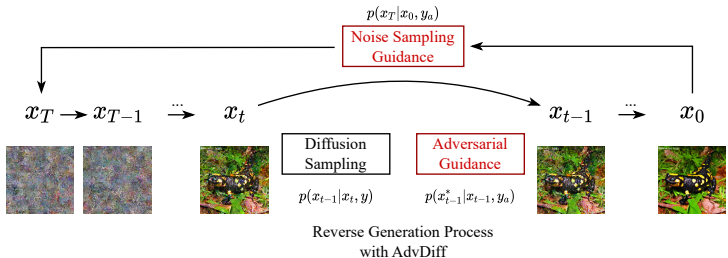


Table of Contents

Evaluation

- Dataset: MNIST and ImageNet
- Comparisons:
  - Perturbation-based attacks: PGD, C&W, and AutoAttack
  - GAN-based unrestricted attacks: U-GAN
  - Diffusion-based unrestricted attacks: DiffAttack, and AdvDiffuser
- Implementation: DDPM model for MNIST and DDIM model for ImageNet
- Because existing diffusion model attacks are all untargeted attacks, we include the untargeted version of AdvDiff for a clear comparison, which is represented by “AdvDiff-Untargeted”.

Evaluation

Method	ASR(%)						Time (s)
	ResNet50			WideResNet50-2			
	Clean	DiffPure	PGD-AT	Clean	DiffPure	PGD-AT	
AutoAttack	95.1	22.2	56.2	94.9	20.6	55.4	0.5
U-SAGAN	99.3	30.5	80.6	98.9	28.6	70.1	10.4
U-BigGAN	96.8	40.1	81.5	96.5	35.5	78.4	11.2
AdvDiffuser	95.4	28.9	90.6	94.6	26.5	88.9	38.6
DiffAttack	92.8	30.6	88.4	90.6	27.6	85.3	28.2
AdvDiff	**99.8**	41.6	92.4	**99.9**	38.5	90.6	**9.2**
AdvDiff-Untargeted	99.5	**75.2**	**94.5**	99.4	**70.5**	**92.6**	9.6

- Our methods outperform all existing methods on white-box attacks and against effective defenses.
- The time efficiency of the proposed methods is also improved.

Evaluation

Method	FID (↓)	LPIPS (↓)	SSIM (↑)	BRISQUE (↓)	TRES (↑)
AutoAttack	26.5	0.72	0.21	34.4	69.8
U-BigGAN	25.4	0.50	0.32	19.4	80.3
AdvDiffuser	26.8	0.21	0.84	18.9	75.6
DiffAttack	20.5	0.15	0.75	22.6	67.8
AdvDiff	**16.2**	**0.03**	**0.96**	**18.1**	**82.1**
AdvDiff-Untargeted	22.8	0.14	0.85	16.2	76.8

- Our methods significantly outperform the perturbation-based method, i.e., AutoAttack in image quality.
- We achieve solid improvements over existing unrestricted adversarial attacks, which validates the need for theoretically supported adversarial guidance.

Evaluation

- Compared with existing diffusion-based adversarial attacks, the perturbation of our method is remarkably smaller.

Table of Contents

Conclusion

- AdvDiff is an effective and interpretable unrestricted adversarial attack using diffusion models.
- AdvDiff can be utilized in any pre-trained diffusion model without re-training or modifying the benign sampling method.
- AdvDiff achieves state-of-the-art performance on white-box attack, transfer attack, and generation quality.

Q&A

Thank you for listening!
Your feedback will be highly appreciated!