Exploring Adversarial Fake Images on Face Manifold

Dongze Li[1,2], Wei Wang[2]*, Hongxing Fan[1,2], Jing Dong[2]

[1] School of Artificial Intelligence, University of Chinese Academy of Sciences [2] Center for Research on Intelligent Perception and Computing, CASIA

Motivation

➢ Problem description

1. GAN-generated images can be efficiently detected by image forensic detectors, thus, for attackers who craft fake images, endowing generated images with anti-forensic property is of great significance.

2. Traditional attack methods can be easily noticed by human eyes because of visible adversarial perturbation added in image space, and result images have low visual quality.

➢ Adversarial Attack

Adding adversarial perturbations on images to fool deep learning models,.

- FGSM [1]

Single step adversarial attack based on the sign of the gradient.

- PGD [2]

Multi step adversarial attack based on projected gradient.

➢ Image Forensics

Binary classification model to determine whether an image is fake or not.

➢ Style-GAN

Proposed in [3], Style-GAN can generate diverse high resolution images. Styles of generated images are controlled by style vector, while randomly initialized noise vectors determine stochastic details of the images. Style-GAN has shown brilliant results on image generation.

➢ Generated images from different methods

Adversarial attacks on face manifold:

➢ Latent Attack

Adding trainable adversarial perturbations on latent vector of Style GAN.

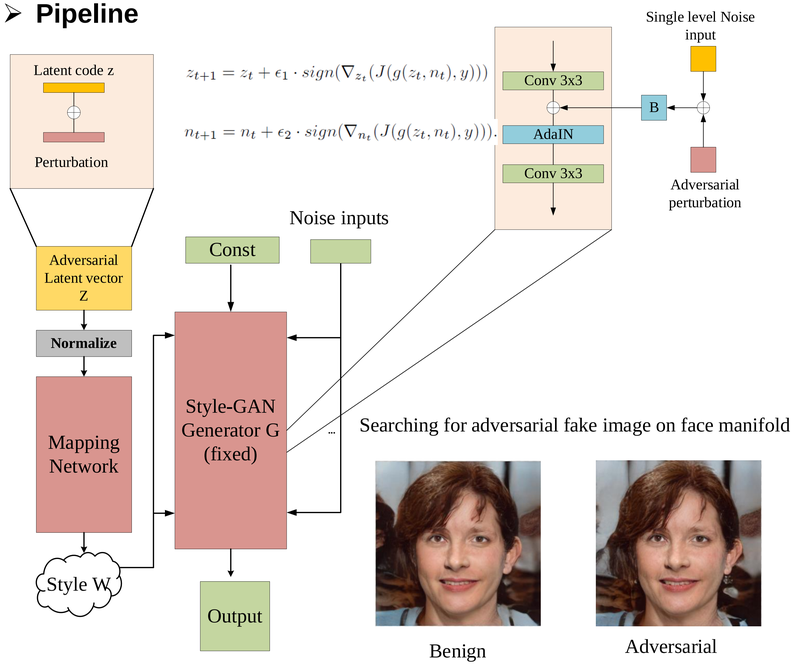
➢ Noise Attack.

Adding trainable adversarial perturbations on noise inputs of different levels.

➢ Ensemble Attack

To make sure our attack can succeed on different models, we craft attack on different models in three ensemble manners: ensemble in logits, ensemble in loss and alternative attacking.

➢ Pipeline



Results

➢ Adversarial attack

Method	Model	
	EfficientNet	Xception
Clean image	97%	93%
PGD $L_{inf}(\epsilon = 0.3)$	**0%**	5%
PGD $L_2(\epsilon = 0.3)$	60%	63%
FGSM L_{inf} $(\epsilon = 0.3)$	13%	5%
Noise(ours)	**0%**	**0%**
Latent(ours)	**0%**	**0%**
Noise and latent(ours)	**0%**	**0%**

Target Model	Method	Test Model	
		EfficientNet	Xception
EfficientNet	FGSM L_{inf}	**0%**	**0%**
	PGD L_{inf}	**0%**	**0%**
	PGD L_2	60%	81%
	Ours	**0%**	50%
Xception	FGSM L_{inf}	**13%**	5%
	PGD L_{inf}	86%	**0%**
	PGD L_2	95%	63%
	Ours	90%	**0%**

➢ Image quality

Metric	Ours	PGD L_{inf}	FGSM L_{inf}
MSE (↓)	0.009	0.027	**0.004**
PSNR (↑)	21.61	15.628	**23.986**
SSIM (↑)	0.891	0.57	**0.926**
LPIPS (↓)	**0.123**	1.084	0.507
User Study	**10000/10000**	0/10000	0/10000

Summary/Conclusion

- A novel method to generate anti-forensic adversarial fake images on face manifold
- Exploration of adversarial attack on GAN latent space.

References

[1] I. J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. In ICLR, 2015.

[2] Madry A, Makelov A, Schmidt L, et al. Towards deep learning models resistant to adversarial attacks. In ICLR, 2018.

[3] Karras T, Laine S, Aila T. A style-based generator architecture for generative adversarial networks[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 4401-4410.