AdvFoolGen: Creating Persistent Troubles for Deep Classifiers

Yuzhen Ding, Nupur Thakur, Baoxin Li
Arizona State University

2021 ICCV OCTOBER 11-17 VIRTUAL

Introduction

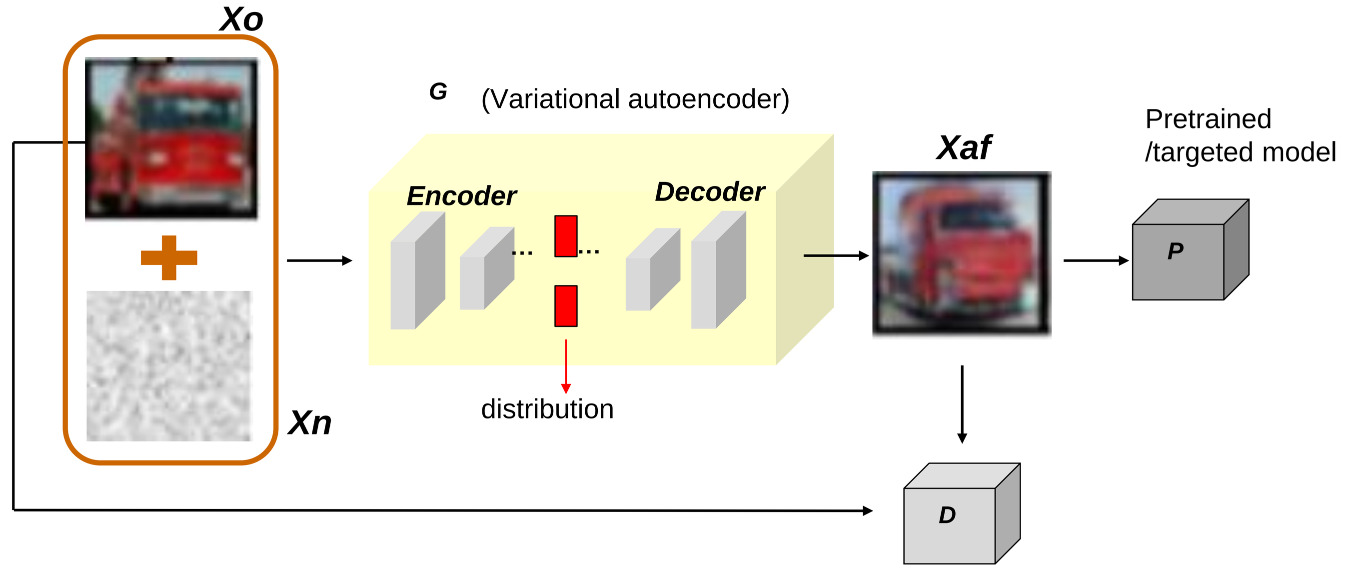
- Deep neural networks are vulnerable to malicious attacks.
- Many defense mechanisms are effective for guarding against typical attacks
- AdvFoolGen, an adversarial attack contributes to understanding the vulnerability of deep networks from a new perspective and may, in turn, help in developing and evaluating new defense mechanisms.

The Technique: AdvFoolGen

- A VAE-GAN-like structure.
- Exploring the latent space where the clean samples lie.
- Generating diverse adversarial/fooling sets by varying the training epochs.

Algorithm 1: AdvFoolGen

Input : Original images X_o, Noise mask X_n, model G, D and P
Output: AdvFool image X_{af}
for *each epoch* $e = 1, 2, \ldots$ **do**
  while *Training* **do**
    for *each batch* $b = 1, 2, \ldots$ **do**
      Generate a noise mask X_n;
      Construct input image as $cat(X_o, X_n)$;
      for $i = 1$ **to** 5 **do**
        Update model D;
      end
      Update model G;
    end
  end
  while *Testing* **do**
    Generate a noise mask X_n;
    Construct input image as $cat(X_o, X_n)$;
    Output X_{af};
    Compute fooling ratio using the predicted label given by P;
  end
end



Experiments

- Initial Fooling ratio:

Attack Algorithm	Initial Fooling Ratio		
	CIFAR10	TinyImageNet	
		Top1	Top5
FGSM	92.82% *	88.55% *	75.18% *
I-FGSM	99% *	100% *	98.86% *
DeepFool	99%	99%	83.77%
C&W	100%	99.12%	90.63%
GAP	82%	94.98%	87.01%
AdvFoolGen	68.5% - 78.36%**	95.41%-97.65%**	90.14%-93.07%**

- Reattack fooling ratio with defense applied:

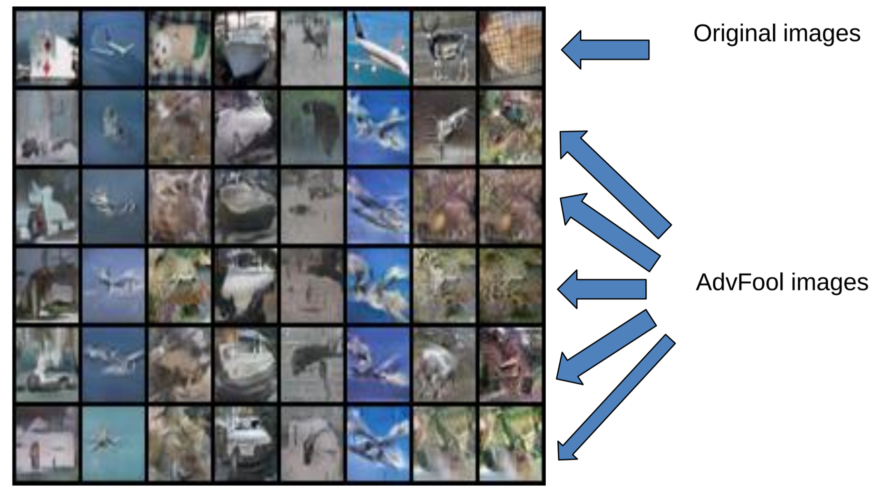
CIFAR10

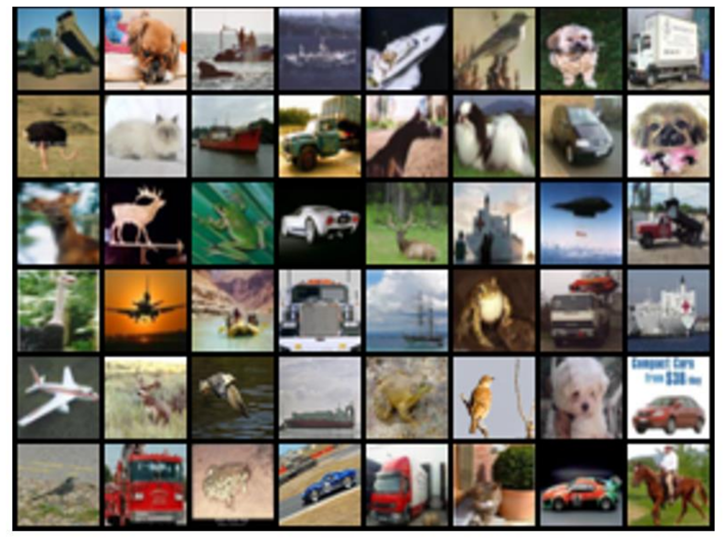
Attack	Retraining*	Adv Training	BDR-3	BDR-8	JPEG
FGSM	9.76%	35.9%	18.21%	16.2%	18.6%
I-FGSM	8.22%	39.3%	12.32%	11.2%	13.1%
DeepFool	9.87%	26.5%	14.55%	14.1%	14.8%
C&W	9.2%	41.25%	12.97%	12.19%	15.67%
GAP	8.91%	9.04%	14.99%	15.09%	19.89%
AdvFoolGen	**27.3%-58.1%**	**59.56%-65.26%**	**37.08%-52.82%**	**24.76%-35.4%**	**24.44%-50.64%**

TinyImageNet

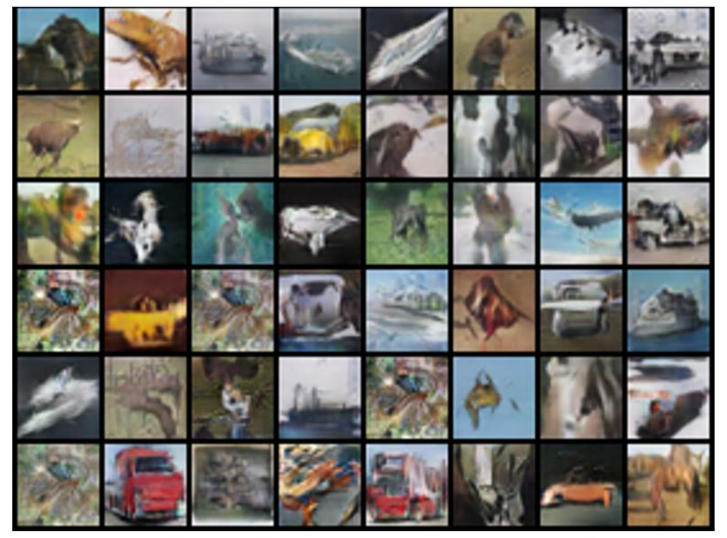
Attack Algorithm	Retraining*	Adversarial Training	BDR-3	JPEG
FGSM	30.8%	49.37%	51.92%	54.18%
I-FGSM	40.5%	48.74%	48.44%	51.15%
DeepFool	29.2%	47.36%	43.02%	47.76%
CW	30.04%	48.61%	46.95%	47.26%
GAP	34.09%	33.76%	33.55%	35.21%
AdvFoolGen**	**43.1%-57.2%**	**54.6%-61.0%**	**40.3%-66.4%**	**42.1%-63.9%**

- Samples generated from different epochs:





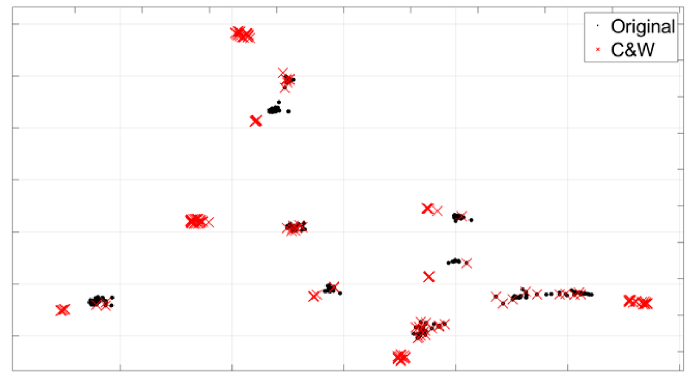
(a)

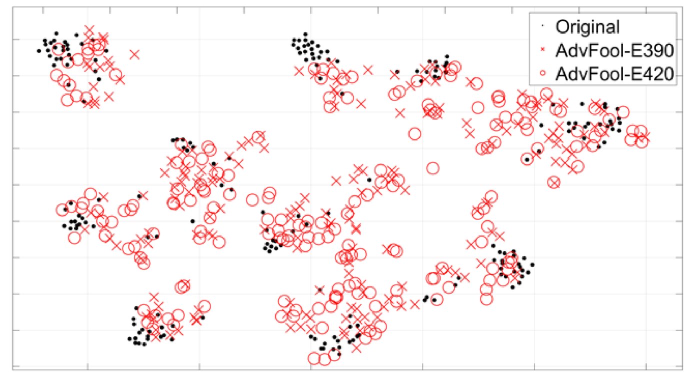


(b)

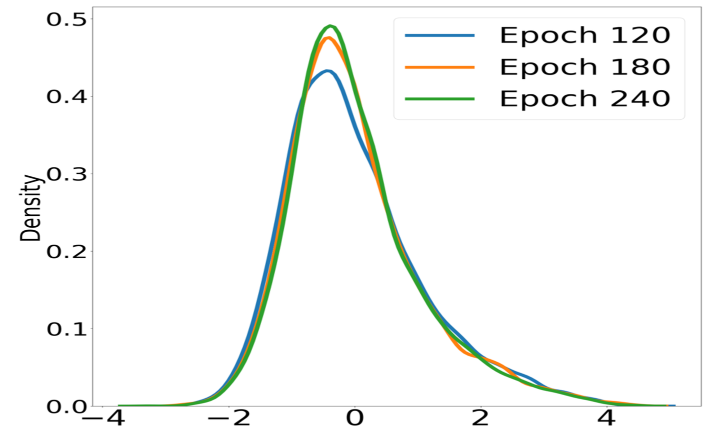
Analysis

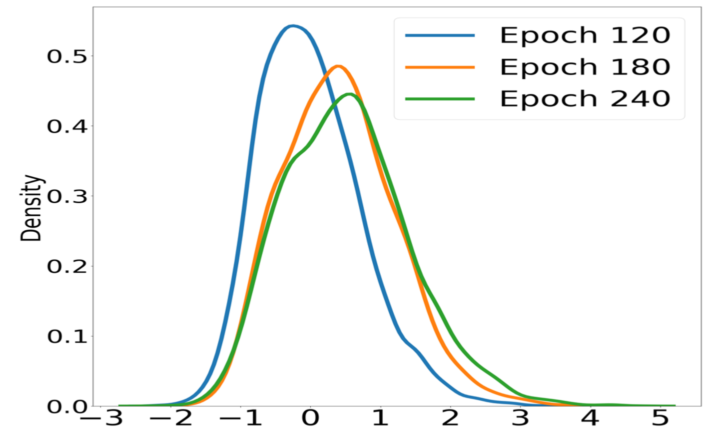
- How advfool samples and original samples behave in the latent space?





- Analytical Study of AdvFool Images





Epoch	Mean		Variance	
	KLD (P‖Q)	KLD (Q‖P)	KLD (P‖Q)	KLD (Q‖P)
120 &180	0.0491	0.0123	1.35	0.299
120 & 240	0.0102	0.0129	2.513	0.313
120 & 330	0.0103	0.0129	2.539	0.314
120 & 360	0.0131	0.0130	2.546	0.314

Summary/Conclusion

- Over-parameterized network leaves backdoors for attackers.
- Simple defenses like retraining can be easily baffled.