Differential Privacy:

Summary:

For inference attacks (attribute, membership), the use of differential privacy/DPSGD provides slight to large amount of privacy improvement, depending in large part on the privacy budget and the training set size of the model. The visual quality of samples from generative models is decreased, but comparable to other generative models.

Attribute inference:

[FW1_42] :

Attribute Inference privacy gain: 2e-4 => 0.037/0.038 (DP on discriminator/generator)

Summary: Slight increase in privacy gain in both generator and discriminator, though GAN is naturally more vulnerable to attribute inference

Membership Inference:

[FW1_42]:

Naïve privacy gain: 0.238 => 0.266/0.245 (DP on discriminator/generator)

Correlation privacy gain: 0.233 => 0.248/0.26

Summary: Privides a slight increase in privacy gain with DP, with DP for discriminator best in naive, and DP for generator best in correlation.

[2]:

FID: Comparable to similar technologies in terms of image quality (FID)

PG-GAN	WGAN-GP	DC-GAN	VAE-GAN	SOTA ref	PGGAN w/DP
14.86	24.26	35.40	53.08	7.40	15.63

AUCROC:

	Full black box	Partial black box	White box
Without DP	0.54	0.58	0.68
With DP	0.53	0.56	0.59

Summary: DP does not inhibit the quality of generated images, and somewhat improves privacy against membership inference in all contexts.

[3]: No measurement on DP against membership inference

[BW1_08]:

Membership inference accuracy: privacy guarantees of σ=1 and σ=5 both remain around 0.5-0.55 for all rounds of testing, while normal GAN increases to around 0.75 after 3000 epochs.

ROC curve: DP trained shows a nearly diagonal line, AUCROC would be around 0.5, with even a small amount of training data. Normal GAN has a worse curve, AUCROC around 0.8 at worst, and 0.6 at best, depending on training set size.

[BW1_09]:

Membership inference accuracy (on 10 classes, random guess yields 0.1):

	Epsilon 1.5 (Good privacy)	Epsilon ~15	Epsilon 28.3
MI accuracy	~0.1	~0.6	0.85

With decrease in epsilon, quality of generated samples also decrease

Summary: No comparison with non-DP, only between different privacy guarantees and their results

[FW1_02]:

Membership inference accuracy:

Dataset	Rand	GAN	privGAN			DPGAN	
			λ = 0.1	λ = 1	λ = 10	ε = 100	ε = 25
MNIST	0.1	0.467	0.144	0.12	0.096	0.098	0.1
f-MNIST	0.1	0.527	0.192	0.192	0.095	0.102	0.099
LFW	0.1	0.724	0.148	0.107	0.086	0.109	0.097
CIFAR-10	0.1	0.723	0.568	0.424	0.154	0.107	0.098

Summary: For all datasets, DP with any λ provides descent, to large decrease In MI accuracy, with λ = 10 for privGAN, and ε = 25 providing equal or less accuracy than random guess. With increasing λ, follows a worsening quality, though not very noticeable for values 0.1, 1, and 10, and In addition, certain classes become overrepresented in the generated data.

[FW1_53]:

Membership inference accuracy:

Dataset	Rand	CDP accuracy			LDP accuracy			VAE-LDP accuracy		
		ε = ~10^8	ε = ~10^5	ε = ~80	ε = 10^4	ε = 1000	ε = 100	σ = 1000	σ = 10	σ = 0.1
LFW	0.5	~0.35	~0.15	~0.1	~0.65	~0.45	~0.1	1	~0.83	0.75
Motion Sense	0.5	~0.55	~0.55	~0.45	NA	NA	NA	NA	~0.45	~0.25

Summary: All types of DP, CDP, LDP and VAE-LDP decreases MI accuracy with variation in σ and ε.

[FW1_43]: No measurement on DP against membership inference

[FW1_25]:

Summary: Advantage degradation is positive for several models when using DP.

[FW1_20]:

Membership inference accuracy:

Dataset	Rand	DP GAN utility accuracy				TableGAN-MCA AUPRC			
		Non-private WGAN	ε = 2	ε = 8	ε = 16	Non-private WGAN	ε = 2	ε = 8	ε = 16
Adult	0.5	0.84	0.83	0.84	0.84	0.68	0.53	0.7	0.7
Lawschool	0.5	0.81	0.78	0.81	0.81	0.38	0.26	0.37	0.38
Compas	0.5	0.65	0.55	0.62	0.63	0.6	NA	0.3	0.6

Summary: Increasing ε increases the MI accuracy.

DPSGD:

Adversarial Examples:

[FW1_15]: Not applicable

Attribute Inference:

[FW1_15]: Not applicable

Evasion:

[FW1_15]: Not applicable

Membership Inference:

[2]:

FID: Comparable to similar technologies in terms of image quality (FID)

PG-GAN	WGAN-GP	DC-GAN	VAE-GAN	SOTA ref	PGGAN w/DP
14.86	24.26	35.40	53.08	7.40	15.63

AUCROC:

	Full black box	Partial black box	White box
Without DP	0.54	0.58	0.68
With DP	0.53	0.56	0.59

Summary: DP does not inhibit the quality of generated images, and somewhat improves privacy against membership inference in all contexts.

[FW1_15]:

Feasible but reduces generated sample quality and increases the training cost

Adversarial Training

Summary:

Adversarial training against adversarial examples provides slight to noticeable improvement in model performance against adversarial attacks, with a slight reduction in performance on clean samples. The adversarial training provides this both with all training samples adversarially trained, and partial adversarial training.

Adversarial Examples:

[17]: Doesn't describe the effectiveness of the attacks, but measures the classification accuracy of the classifier using the output from the autoencoder.

[18]:

Dataset	Defense	δ max of Linf attacks			
		0	0.02	0.04	0.08
CIFAR10	Adv training	81.45%	69.15%	53.74%	23.58%
	Rob-GAN	81.1%	70.41%	57.43%	30.25%
		0	0.01	0.02	0.03
ImageNet	Adv training	20.05%	18.3%	12.52%	8.32%
	Rob-GAN	32.4%	25.2%	19.1%	13.7%

Summary: This papers defense is more robust than regular adversarial training, in most cases, for both datasets. No comparison to non-defensive GAN.

[21]: No measurement of the defense

[FW1_10]:

Clustering accuracy of adversarially trained reaches ~0.87 for both clean and perturbed samples

[FW1_19]:

"Robustness" of the models:

Dataset	ε	Clean Train	Adversarial Train	Hybrid Train
CIFAR-10	0	3.4	4.7	3.6
	1	6.3	4.9	4.7
	2	14	5.0	5.0

	4	320	5.3	5.3
	8	2*10^6	5.8	5.9
LSUN-Bedrooms	0	2.4	4.4	2.9
	1	5.5	4.5	4.7
	2	8.5	4.7	4.6
	4	15.2	5.0	5.0
	8	27.0	5.5	5.6
	16	35.8	6.6	6.6
	32	36.4	7.7	8.1

Summary: Both fully adversarially trained, and half-clean-half-perturbed adversarially trained models shows a much higher degree of "robustness", compared to non adversarially trained.

[FW1_37]: No measurement of the defense

[FW2_10]:

Accuracy of model:

Training of linear head	Norm	Radius	ImageNet 100%		ImageNet 10%		ImageNet 1%	
			Clean	Robust	Clean	Robust	Clean	Robust
Robust	Linf	$\epsilon = 4/255$	65.14%	45.44%	62.39%	41.58%	47.07%	31.57%
Non-robust			65.27%	43.83%	62.40%	41.06%	47.64%	31.99%
Robust	L2	$\epsilon = 128/255$	69.64%	65.48%	66.39%	61.90%	55.06%	51.00%

Summary: Adversarial training provides some benefit under Linf norm, but much more (though not certain how it is compared to non-robust) in L2 norm.

Poisoning:

[FW1_37]: No measurement of the defense

Model Extraction:

[FW1_15]: Not applicable

Model Inversion:

[FW1_15]: Not applicable

Poisoning:

[FW1_15]: Not applicable

Gradient Masking:

Summary:

Gradient masking against adversarial examples shows improvement of model performance, in particular in white-box settings.

Adversarial Examples:

[17]:
Using gradient masking methods with objectives “temperature T of 2.5” and Guided Complement Entropy (GCE).

Shows increase in classifier accuracy in white-box setting, and marginal increase in black-box setting. GCE appears a bit more robust than temperature T.

Backdoor:

[FW1_13]: Does not show results for this.

Game Theory:

Evasion:

[12]: Does not show results

Poisoning:

[12]: Does not show results

Fine Pruning:

Adversarial Examples:

[FW1_15]: Not applicable

Attribute inference:

[FW1_15]: Not applicable

Evasion:

[FW1_15]: Not applicable

Membership inference:

[FW1_15]: Not applicable

Model extraction:

[FW1_15]: Not applicable

Model Inversion:

[FW1_15]: Not applicable

Poisoning:

[FW1_15]: Does not work for generative models, and decreases sample quality

Bootstrapping:

Adversarial Examples:

[FW2_10]:

Comparing accuracy between adversarial training and a bootstrapped model

Robust bootstrapping provides a lot of improvement on adversarial samples versus non-robust, though non-robust is marginally better at clean samples (under L2 norm constraint).

At most datasets/number of labeled samples, the bootstrap performs better than adversarial training under L2 norm constraint, and better, or marginally worse, than adversarial training under Linf constraint.

Weight Normalization:

Adversarial Examples:

[FW1_15]: Not applicable

Attribute Inference:

[FW1_15]: Not applicable

Evasion:

[FW1_15]: Not applicable

Membership inference:

[FW1_15]:

Can improve generalizability, but can lead to instability, where generator or discriminator can outperform the other.

Model Extraction:

[FW1_15]: Not applicable

Model Inversion:

[FW1_15]: Not applicable

Poisoning:

[FW1_15]: Not applicable

Dropout:

Adversarial Examples:

[FW1_15]: Not applicable

Attribute Inference:

[FW1_15]: Not applicable

Evasion:

[FW1_15]: Not applicable

Membership inference:

[FW1_15]:

Prevents overfitting, but reduces quality, in turn requiring larger epochs to train to a satisfactory level.

Model Extraction:

[FW1_15]: Not applicable

Model Inversion:

[FW1_15]: Not applicable

Poisoning:

[FW1_15]: Not applicable