Self-Diagnosing GAN: Diagnosing Underrepresented Samples in Generative Adversarial Networks

Jinhee Lee*, Haeri Kim*, Youngkyu Hong*, Hye Won Chung

*: equal contribution
Korea Advanced Institute of Science and Technology (KAIST)

NeurIPS 2021

Motivation and Our Goal

GANs are good at generating high-quality realistic images, but...

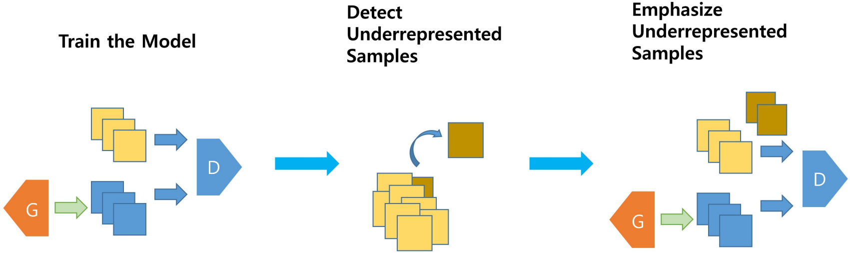
1. Often fail to learn **sparse regions** of data manifold, e.g. having poor modeling for samples with **minor features** [Karras et al. 2018; DeVries et al. 2020; Yu et al. 2020]
2. Suffer from **mode collapse** [Lin et al. 2018]

$\Rightarrow$ There exist underrepresented samples!

Goal: Improve **diversity** in sample generation while not degrading the overall **quality**

Overview of Our Strategy

We design methods to **detect** and **emphasize** underrepresented samples in training of GANs



Challenges in Detecting Underrepresented Samples

- To diagnose underrepresented samples, try to measure

$$\text{Log density ratio: } \log \frac{p_{\text{data}}(x)}{p_g(x)}$$

- But unknown data distribution p_{data} and implicit model distribution p_g
- **Idea**: approximate the log density ratio by using the discriminator output $D(x)$,

$$\text{LDR}(x) := \log \frac{D(x)}{1 - D(x)} \approx \log \frac{p_{\text{data}}(x)}{p_g(x)}$$

- How can we do this?

Log Density Ratio (LDR) Estimate

Definition: LDR Estimate

$$\mathrm{LDR}(x) := \log \frac{D(x)}{1 - D(x)} \approx \log \frac{p_{\text{data}}(x)}{p_g(x)}$$

- GAN solves the min-max optimization:

$$\min_G \max_D \{\mathbb{E}_{x \sim p_{\text{data}}}[\log D(x)] + \mathbb{E}_{z \sim p_z}[\log(1 - D(G(z)))]\}$$

- **Optimal discriminator**:

$$D^*(x) = \frac{p_{\text{data}}(x)}{p_{\text{data}}(x) + p_g(x)}$$

$$\text{When } D(x) = D^*(x) \quad \Rightarrow \quad \mathrm{LDR}(x) = \log \tfrac{p_{\text{data}}(x)}{p_g(x)}$$

What LDR Might Tell?

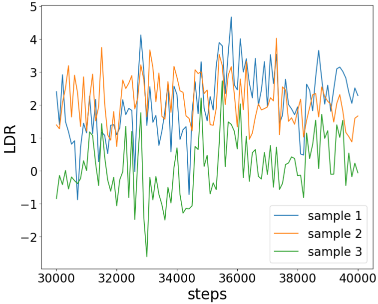
LDR

$$\mathrm{LDR}(x) := \log \frac{D(x)}{1 - D(x)} \approx \log \frac{p_{\mathrm{data}}(x)}{p_g(x)}$$

- $\mathrm{LDR}(x) > 0$: the data x is underrepresented, i.e., $p_{\mathrm{data}}(x) > p_g(x)$
- $\mathrm{LDR}(x) < 0$: the data x is overrepresented, i.e., $p_{\mathrm{data}}(x) < p_g(x)$

But LDR is Unstable

LDR values are unstable during training.



- Hard to diagnose GAN training from the LDR at a particular step
- **Idea**: use the **mean** and **variance** of the LDRs over multiple steps

New Measures: LDRM and LDRV

- **LDRM** (LDR Mean)

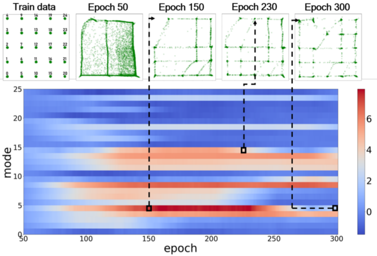
$$\mathrm{LDRM}(x;T) = \frac{1}{|T|}\sum_{k\in T}\mathrm{LDR}(x)_k,$$

- **LDRV** (LDR Variance)

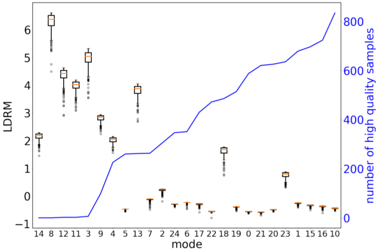
$$\mathrm{LDRV}(x;T) = \frac{1}{|T|-1}\sum_{k\in T}[\mathrm{LDR}(x)_k - \mathrm{LDRM}(x;T)]^2$$

at each sample x across the training steps $T = \{t_s, ..., t_e\}$

Training dynamics of 25 Gaussian and LDRM for training samples



(a) LDRM and generated samples



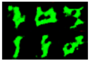
(b) LDRM distribution and the number of high-quality samples over modes

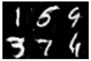
- Modes with high LDRM do not appear in the generated samples
- **LDRM is effective in detecting missing modes**

LDRV is Effective in Detecting Minor Features

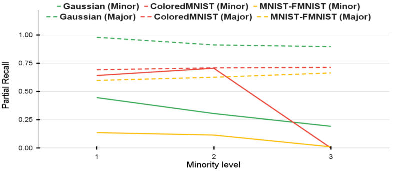
Generated sample quality for major-minor attributes


(a) Major


(b) Minor


(c) Major


(d) Minor



(e) Partial Recall of major/minor groups vs. minority level

LDRV is higher for minor samples

Group	Gaussian (σ=3.0)	Colored MNIST	MNIST-FMNIST
Major	0.001	0.077	0.082
Minor	0.098	0.186	0.115

- GANs have poor modeling for minor samples.
- **LDRV is effective in detecting minor features**

Theoretical Insights on LDRV

- Viewing the discriminator as the **logistic regression model**,

$$\mathrm{LDRV}(x_i) \approx \mathrm{var}\left(\log(D(x_i;\theta)/(1-D(x_i;\theta)))\right) \approx \phi_i^T S_n \phi_i.$$

with the feature vector ϕ_i of each data x_i and covariance matrix

$$S_n = \left(\sum_{i=1}^{n} D(x_i;\theta)(1-D(x_i;\theta))\phi_i\phi_i^T + \frac{1}{s_0}I\right)^{-1}.$$

- Minor feature vector, which has a small component on the eigenspace formed by the majority of $\{\phi_i\}$, tends to have a higher LDRV

Discrepancy Score

Definition: Discrepancy score

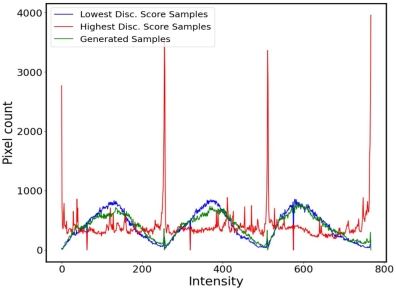
$$s(x_i; T) := \mathrm{LDRM}(x_i; T) + k\sqrt{\mathrm{LDRV}(x_i; T)},$$

Discrepancy score reflects the underrepresentedness of each sample

(a) Images with lowest disc. score

(b) Images with highest disc. score

(c) Generated samples



(d) Histogram of pixel count over intensity level

Our Algorithm: Emphasizing Underrepresented Samples

1. **Phase 1 - Train and Diagnose**
   Train GAN and evaluate the discrepancy score $s(x_i)$ for each data instance x_i

2. **Phase 2 - Score-Based Weighted Sampling**
   Encourage GAN to learn underrepresented regions of data manifold through score-based weighted sampling: set the minibatch sampling frequency $P_s(i)$ proportional to $s(x_i)$

$$\mathcal{D}_B = \{x^{(j)} : x^{(j)} = x_i \text{ where } i \sim P_s(i) \propto s(x_i) \text{ for } j = 1, \ldots, B\}$$

3. **Phase 3 - DRS**
   After GAN training, correct the model distribution $p_g(x)$ by rejection sampling

Experimental Results: Improving the Overall Performance

- Our method is effective in **improving the overall quality**, measured by both fidelity and diversity combined

Table 2: Comparison of diverse sampling/weighting methods for CIFAR-10/CelebA image generation.

Dataset	CIFAR-10				CelebA					
Methods	SNGAN		SSGAN		SNGAN			SSGAN		
	FID ↓	IS ↑	FID ↓	IS ↑	FID ↓	P ↑	R ↑	FID ↓	P ↑	R ↑
Vanilla	26.90	7.36	22.01	7.65	7.12	0.68	0.44	7.19	0.68	0.44
DRS [3]	24.54	7.57	20.51	7.77	7.04	0.68	0.44	7.08	0.68	0.45
GOLD [32]	28.86	7.21	21.90	7.57	7.31	**0.69**	0.44	7.46	0.68	0.43
GOLD + DRS	24.65	7.53	19.36	7.79	6.97	0.68	0.44	7.15	0.67	0.45
Top-k [37]	24.45	7.60	20.01	7.78	7.35	0.67	0.44	7.23	0.67	0.45
Top-k + DRS	23.92	7.70	20.09	7.88	7.35	0.68	0.44	7.16	**0.68**	0.45
Dia-GAN	**19.66**	**7.95**	**16.31**	**8.14**	**6.70**	0.64	**0.48**	**6.88**	0.66	**0.46**

Experimental Results: Scaling to a Larger Model

- Our method is scalable to **large state-of-the-art GANs** and **high resolution images**
- Our method may be **extended to hinge-loss variant** of GANs

Table 3: StyleGAN2 on FFHQ 256x256.

	FID ↓	P ↑	R ↑
StyleGAN2	14.07	**0.72**	0.27
GOLD	15.33	0.69	0.29
Dia-StyleGAN2	**11.89**	0.69	**0.30**

Table 4: HingeGAN on CIFAR-10 and CelebA.

	CIFAR-10		CelebA
	FID ↓	IS ↑	FID ↓
HingeGAN	21.99	7.67	6.66
Dia-HingeGAN	**18.74**	**8.02**	**5.98**

Experimental Results: Minor Feature Enhancement

- Our **LDRV score effectively detects minor samples** in real dataset
- Improves the **occurrence rate** and **partial recall** of minor attributes

Table 6: CelebA minor attribute analysis. Averaged LDRV and averaged discrepancy score of CelebA samples with (W/) or without (W/O) minor attributes. O stands for the occurrence of minor attributes among the generated samples in percentage (%) and R stands for the Partial Recall.

	Score				Method			
	LDRV		Discrepancy		Vanilla		**Dia-GAN**	
	W/	W/O	W/	W/O	O ↑	R ↑	O ↑	R ↑
Bald (2.244%)	**0.271**	0.184	**2.938**	2.221	0.678	0.353	**0.836**	**0.393**
Double Chin (4.669%)	**0.219**	0.184	**2.525**	2.224	0.440	0.411	**0.522**	**0.461**
Eyeglasses (6.512%)	**0.254**	0.181	**2.783**	2.200	3.300	0.400	**4.053**	**0.449**
Gray Hair (4.195%)	**0.211**	0.185	**2.450**	2.228	2.273	0.402	**2.369**	**0.436**
Mustache (4.155%)	**0.242**	0.183	**2.699**	2.218	0.157	0.391	**0.228**	**0.433**
Pale Skin (4.295%)	**0.190**	0.186	**2.240**	2.238	0.346	0.380	**0.453**	**0.427**
Wearing Hat (4.846%)	**0.357**	0.177	**3.651**	2.164	2.307	0.380	**3.595**	**0.408**

Summary: Self-Diagnosing GAN

- Provided two measures to detect underrepresented samples
  - LDRM: effective in detecting missing modes
  - LDRV: effective in detecting underrepresented minor features
- Proposed an algorithm enhancing the diversity in sample generation, with special care for minor attributes
- Our method provides a way to diagnose GAN training and can be extended for further improvement of GANs