Reg-DGM

Deep Generative Modeling on Limited Data with Regularization by Nontransferable Pre-trained Models

Yong Zhong[1,2], Hongtao Liu[1,2], Xiaodong Liu[1,2], Fan Bao[3], Weiran Shen[1,2], Chongxuan Li[1,2]

[1]Gaoling School of AI, Renmin University of China, Beijing, China
[2]Beijing Key Lab of Big Data Management and Analysis Methods, Beijing, China
[3]Department of Computer Science Technology, Tsinghua University, Beijing, China

June 9, 2023

Outline

Research Background

Introduction

- GANs produce poor samples with limited data.
- The problem is shared by other DGMs.

Related Work

- Data augmentations.
- Designing new losses.
- Transferring a pre-trained DGM.

Method

Motivation

Inspired by the bias-variance dilemma, we propose a complementary framework **Reg-DGM**, which leverages a pre-trained model to reduce the variance of training a DGM with limited data.

Our Method

Let x denote the real or fake sample, $p_d(x)$ denote the distribution of real data, $p_g(x)$ denote the generator's distribution, $\mathbb{D}(\cdot||\cdot)$ denote a proper statistical divergence, and $\mathcal{R}_f(x) : \mathcal{X} \to \mathbb{R}$ denote the loss from the per-trained model f, we can define our objective loss function:

$$\min_{p_g(x)} \mathbb{D}(p_d(x)||p_g(x)) + \lambda \mathbb{E}_{x \sim p_g(x)}[\mathcal{R}_f(x)], \quad (1)$$

where $\lambda \geq 0$ is a hyperparameter to control the relative weight of the two terms.

Method

A Prototypical Gaussian-fitting Example

The data distribution is a (univariate) Gaussian $p_d(x) = \mathcal{N}(x|\mu^*, \sigma^2)$, where σ^2 is known and μ^* is the parameter to be estimated. A training sample $\mathcal{S} = \{x_i\}_{i=1}^m$ is drawn i.i.d. according to $p_d(x)$. The hypothesis class for p_g is $\mathcal{H} = \{\mathcal{N}(x|\mu, \sigma^2) \mid \mu \in \mathbb{R}\}$. The regularization term in Eq. (1) is $\mathcal{E}_f(x) := -\log \mathcal{N}(\hat{\mu}_{\mathrm{PRE}}, \sigma^2)$, i.e., $p_f(x) = \mathcal{N}(x|\hat{\mu}_{\mathrm{PRE}}, \sigma^2)$.

Proposition 2.2

Let $\beta = \frac{\lambda}{\lambda+1}$ be the normalized weight of the regularization term. In the Gaussian-fitting example, if $\max\left\{\frac{\sigma^2 - m(\hat{\mu}_{\mathrm{PRE}} - \mu^*)^2}{\sigma^2 + m(\hat{\mu}_{\mathrm{PRE}} - \mu^*)^2}, 0\right\} < \beta < \min\left\{\frac{2\sigma^2}{\sigma^2 + m(\hat{\mu}_{\mathrm{PRE}} - \mu^*)^2}, 1\right\}$, then the following inequalities holds:

$$\mathrm{MSE}[\hat{\mu}_{\mathrm{REG}}] < \min\{\mathrm{MSE}[\hat{\mu}_{\mathrm{MLE}}], \mathrm{MSE}[\hat{\mu}_{\mathrm{PRE}}]\}. \tag{2}$$

Convergence Analyses

Analyses in the Non-parametric Setting

Theorem 3.1

Under mild regularity conditions in Assumption A.1, for any $\lambda > 0$, there exists a unique global minimum of the problem in Eq. (1) with the KL divergence. Furthermore, the global minimum is in the form of $p_g^*(x) = \frac{p_d(x)}{\alpha^* + \lambda \mathcal{E}_f(x)}$, where $\alpha^* \in \mathbb{R}$.

Theorem 3.2

Under mild regularity conditions in Assumption A.1, for any $\lambda > 0$, there exists a unique global minimum of the problem in Eq. (1) with the JS divergence. Furthermore, the global minimum is in the form of $p_g^*(x) = \frac{p_d(x)}{e^{\alpha^* + \lambda \mathcal{E}_f(x)} - 1}$, where $\alpha^* \in \mathbb{R}$.

Convergence Analyses

Analyses in the Parametric Settings

Theorem 3.3 (Convergence of Reg-DGM (informal))

Under standard and verifiable smoothness assumptions, with a high probability, Reg-DGM with a sufficiently wide ReLU CNN converges to a global optimum of Eq. (1) trained by GD and converges to a local minimum trained by SGD.

Implementation

Base Model

StyleGAN2, adaptive discriminator augmentation (ADA), and adaptive pseudo augmentation (APA).

Pre-trained Model

ResNet, CLIP image encode, and FaceNet.

Energy Function

The energy function is defined by the expected mean squared error between the features of a generated sample and a training sample as follows:

$$\mathcal{E}_f(x) := \mathbb{E}_{x' \sim p_d} \left[\frac{1}{d} ||f(x) - f(x')||_2^2 \right]. \tag{3}$$

Experiments

Benchmark Results with Limited Data

Table 1: Median FID ↓ on FFHQ and LSUN CAT and mean FID ↓ on CIFAR-10. † and ‡ indicate the results are taken from the references and Karras et al. (2020a) respectively. Otherwise, the results are reproduced by us upon the official implementation (Karras et al., 2020a; Jiang et al., 2021).

Method	FFHQ		LSUN CAT		CIFAR-10
	1k	5k	1k	5k	50k
Transfer (Wang et al., 2018)	21.42	12.34			
Freeze-D (Mo et al., 2020)	19.77	12.69			
DA† (Zhao et al., 2020a)	25.66	10.45	42.26	16.11	8.49
InsGen† (Yang et al., 2021)	19.58				
GenCo† (Cui et al., 2021)	65.31	27.96	140.08	40.79	8.83 ± 0.04
DA + GenCo† (Cui et al., 2021)					6.57 ± 0.01
ADA + bCR‡ (Zhao et al., 2020b)	22.61	10.58	38.82	16.80	
R_{LC} † (Tseng et al., 2021)	63.16	23.83			8.31 ± 0.05
ADA + R_{LC}† (Tseng et al., 2021)	21.7				$\mathbf{2.47 \pm 0.01}$
APA† (Jiang et al., 2021)	45.19	13.25			
StyleGAN2 (Karras et al., 2020b)	103.66	52.71	186.55	115.16	7.16 ± 0.12
Reg-StyleGAN2 (**ours**)	75.99	37.77	107.02	63.10	6.56 ± 0.14
ADA (Karras et al., 2020a)	22.26	12.64	41.81	16.76	3.07 ± 0.08
Reg-ADA (**ours**)	20.05	11.95	36.17	15.91	2.95 ± 0.05
ADA + APA (Jiang et al., 2021)	19.71	8.84	24.09	11.79	2.64 ± 0.08
Reg-ADA-APA (**ours**)	**17.88**	**8.02**	**21.88**	**11.27**	2.58 ± 0.04

Experiments

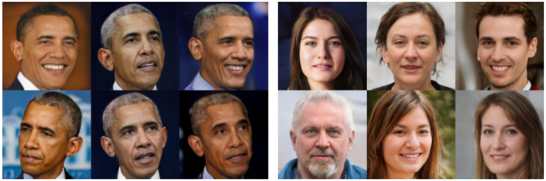
Ablation of Pre-trained Models and Pre-training Datasets

Table 2: Median FID $\downarrow$ and the corresponding KID$\times 10^3$ $\downarrow$ using a pre-trained CLIP or FaceNet.

Method	CLIP				FaceNet	
	FFHQ-5k		LSUN CAT-5k		FFHQ-5k	
	FID	KID	FID	KID	FID	KID
StyleGAN2 (Karras et al., 2020b)	52.71	39.52	115.16	100.57	52.71	39.52
Reg-StyleGAN2(**ours**)	40.98	27.56	42.04	26.21	38.80	23.38
ADA (Karras et al., 2020a)	12.64	5.17	16.76	8.13	12.64	5.17
Reg-ADA(**ours**)	11.09	3.91	14.15	6.72	11.37	4.01
ADA+APA (Jiang et al., 2021)	8.84	2.76	11.79	4.86	8.84	2.76
Reg-ADA-APA(**ours**)	**8.18**	**2.26**	**10.47**	**4.68**	**8.21**	**2.37**

Experiments

Qualitative Result



(a) 100-shot Obama (FID 39.53) (b) FFHQ-5k (FID 11.69)

Figure 3: Samples from the Reg-ADA, truncated ($\psi = 0.7$) as in prior work (Karras et al., 2020a).

Thanks for your attention.