Diagnosing and Enhancing VAE models (ICLR 2019)

Bin Dai, David Wipf

Presenter:
Minhao Jiang (minhaoj2)

Note: A shorter version of this paper is accepted by ICLR 2019 conference proceedings.

Introduction

- It is commonly believed that Gaussian encoder/decoder assumptions reduce the effectiveness of VAEs in generating realistic samples.
- This paper rigorously analyzes that reaching the global optimum does not guarantee that if VAE model can learn the true distribution of data, i.e., there could exist **alternative solutions** that both reach the global optimum and yet do not assign the same probability measure as ground-truth probability distribution.
- The paper proposed a two-stage remedy model, i.e., a **two-stage VAE model** to enhance the original VAE so that any globally minimizing solution is uniquely matched to the ground-truth distribution.

Problem Definition:

- The starting point is the desire to learn a probabilistic generative model of observable variables $x \in \mathcal{X}$ where $\mathcal{X}$ is a r-dimensional manifold embedded in $\mathbb{R}^d$
  - When $r = d$, this assumption places no restrictions on the distribution.
  - When $r \ll d$, this situation is very applicable in the utility of generative models.
- Denote a ground-truth probability measure on $\mathcal{X}$ as μ_{gt} where $\int_{\mathcal{X}} \mu_{gt} dx = 1$
- The canonical VAE attempts to approximate this ground-truth measure using parameterized density $p_\theta(x)$ where $p_\theta(x) = \int p_\theta(x|z)p(z)dz$, $z \in \mathbb{R}^\kappa$ with $\kappa \approx r$ and $p(z) = \mathcal{N}(z|0, I)$

- ELBO:

$$\begin{aligned}\mathcal{L}_{\theta,\phi}(x) &= -\mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x,z) - \log q_\phi(z|x)] \\ &= \mathbb{KL}(q_\phi(z|x)||p_\theta(z)) + \mathbb{E}_{q_\phi(z|x)}[-\log p_\theta(x,z)]\end{aligned}$$

- Another form:

$$\mathcal{L}_{\theta,\phi}(x) = \int_{\mathcal{X}} \{-\log p_\theta(x) + \mathbb{KL}[q_\phi(z|x)||p_\theta(z|x)]\}\mu_{gt}\,dx \geq \int_{\mathcal{X}} -\log p_\theta(x)\mu_{gt}\,dx$$

$$\mathcal{L}_{\theta,\phi}(x) = \int_{\mathcal{X}} \{-\mathbb{E}_{q_\phi(z|x)}[\log p_\theta(z|x)] + \mathbb{KL}[q_\phi(z|x)||p(z)]\}\mu_{gt}\,dx$$

- In principle, $q_\phi(z|x)$ and $p_\theta(x|z)$ can be arbitrary distributions. In the practical implementation, a commonly adopted distributional assumption is that both distribution are Gaussian, which was previously considered as a limitation of VAE.

Diagnosing the Non-uniqueness

Ideas: Even with the stated Gaussian distributions, there exist parameters θ, ϕ that can simultaneously:

1. Globally optimize the VAE object
2. Recover the ground-truth probability measure in a certain sense

Definition 1

A κ-simple VAE is defined as a VAE model with dim[z] $= \kappa$ latent dimensions, the Gaussian encoder $q_\phi(z|X) = \mathcal{N}(z|\mu_z, \Sigma_z)$ and the Gaussian decoder $p_\theta(x|z) = \mathcal{N}(x|\mu_x, \Sigma_x)$

With these definitions, the paper shows that a κ-simple VAE with $\kappa \geq r$ can achieve the above optimality criteria. We will consider this from the simpler case where $r = d$ followed by the extended scenario with $r < d$.

Diagnosing the Non-uniqueness ($r = d$)

Assuming $p_{gt}(x) = \mu_{gt}(dx)/dx$ exists everywhere in $\mathbb{R}^d$, the minimal possible value of negative log-likelihood will necessarily occur if

$$\mathbb{KL}[q_\phi(z|x)||p_\theta(z|x)] = 0 \text{ and } p_\theta(x) = p_{gt}(x) \text{ almost everywhere}$$

Theorem 2

Suppose that $r = d$ and there exists a density $p_{gt}(x)$ associated with the ground-truth measure μ_{gt} that is nonzero everywhere on $\mathbb{R}^d$. Then for any $\kappa \geq r$, there is a sequence of κ-simple VAE model parameters $\{\theta_t^\star, \phi_t^\star\}$ such that

$$\lim_{t\to\infty} \mathbb{KL}[q_{\phi_t^\star}(z|x)||p_{\theta_t^\star}(z|x)] = 0 \text{ and } \lim_{t\to\infty} p_{\theta_t^\star}(x) = p_{gt}(x) \text{ almost everywhere}$$

The theorem implies that as long as latent dimension is sufficiently large (i.e., $\kappa \geq r$), the optimal ground-truth probability measure can be recovered, Gaussian assumptions or not.

Diagnosing the Non-uniqueness ($r < d$)

- When both $q_\phi(z|x)$ and $p_\theta(x|z)$ are arbitrary/unconstrained, then $\inf_{\phi,\theta} \mathcal{L}(\theta, \phi) = -\infty$ by forcing $q_\phi(z|x) = p_\theta(z|x)$.
- To show that this does not need to happen, define a manifold density $\tilde{p}_{gt}(x)$ as the probability density of μ_{gt} with respect to the volume measure of the manifold $\mathcal{X}$. If $d = r$ then this volume is the standard Lebesgue measure in $\mathbb{R}^d$ and $\tilde{p}_{gt}(x) = p_{gt}(x)$

Theorem 3

Assume $r < d$ and that there exists a manifold density $\tilde{p}_{gt}(x)$ associated with the ground-truth measure μ_{gt} that is nonzero everywhere on $\mathcal{X}$. Then for any $\kappa \geq r$, there is a sequence of κ-simple VAE model parameters $\{\theta_t^\star, \phi_t^\star\}$ such that

1. $$\lim_{t\to\infty} \mathbb{KL}[q_{\phi_t^\star}(z|x)||p_{\theta_t^\star}(z|x)] = 0 \text{ and } \lim_{t\to\infty} \int_{\mathcal{X}} -\log p_{\theta_t^\star}(x)\mu_{gt}dx = -\infty$$
2. $$\lim_{t\to\infty} \int_{\mathcal{X}\in A} p_{\theta_t^\star}(x)dx = \mu_{gt}(A \cup \mathcal{X})$$

   for all measurable sets $A \subseteq \mathbb{R}^d$ with $\mu_{gt}(\partial A \cup \mathcal{X}) = 0$ where ∂A is the boundary of A.

Diagnosing the Non-uniqueness ($r < d$)

Implications of Theorem 3:

- From (1), the VAE Gaussian assumptions do not prevent minimization of $\mathcal{L}(\theta, \phi)$ from converging to minus infinity.
- From (2), there exists solutions that assign a probability mass to most all measurable subsets of $\mathbb{R}^d$ that is distinguishable from the ground-truth measure.
- In $r = d$ situation, the theorem necessitates that the ground-truth probability measure has been recovered almost everywhere.
- In $r < d$ situation, we have not ruled out the possibility that a different set of parameters $\{\theta, \phi\}$ can push the lost to $-\infty$ and not achieve (2) ,i.e., the VAE can reach the lower bound of negative log-likelihood but fail to closely approximate μ_{gt}.

Optimal Solution

Necessary Conditions for VAE optima:

Theorem 4

Let $\{\theta^\star_\gamma, \phi^\star_\gamma\}$ denote an optimal κ-simple VAE solution (with $\kappa \geq r$) where the decoder variance γ is fixed. Moreover, we assume that μ_{gt} is not a Gaussian distribution when $d = r$. Then for any $\gamma > 0$, there exists a $\gamma' < \gamma$ such that $\mathcal{L}(\theta^\star_{\gamma'}, \phi^\star_{\gamma'}) < \mathcal{L}(\theta^\star_\gamma, \phi^\star_\gamma)$

The theorem implies that if γ is not constrained, it must be that $\gamma \to 0$ if we wish to minimize the VAE objective. While in existing practical VAE applications, it is standard to fix $\gamma \approx 1$ during training.

Optimal Solution

Theorem 5

Applying the same conditions and definitions in Theorem 4, then for all x drawn from μ_{gt}, we also have that

$$\lim_{\gamma \to 0} f_{\mu_x}[f_{\mu_z}(x; \phi^\star_\gamma) + f_{S_z}(x; \theta\gamma^\star)\epsilon; \phi^\star_\gamma] = \lim_{\gamma \to 0} f_{\mu_x}[f_{\mu_z}(x; \phi^\star_\gamma); \theta^\star_\gamma] = x, \forall \epsilon \in \mathbb{R}^\kappa$$

- With this theorem, it indicates that any x $\in \mathcal{X}$ will be perfectly reconstructed by the VAE model at globally optimal solutions.
- Adding dimensions to latent dimension cannot improve the value of the VAE data term in meaningful way.
- If VAE model parameters have learned a near optimal mapping onto $\mathcal{X}$ using $\gamma \approx 0$, then the VAE cost will scale as $(d - r) \log \gamma$ regardless of μ_{gt}.

Two-Stage VAE model

The above analysis suggests the following two-stage remedy:

1. Given n observed samples $\{x^{(i)}\}_{i=1}^n$, train a κ-simple VAE, with $\kappa \geq r$, to estimate the unknown r-dimensional ground-truth manifold $\mathcal{X}$ embedded in $\mathcal{R}^d$ using a minimal number of active latent dimensions. Generate latent samples $\{z^{(i)}\}_{i=1}^n$ via $z_{(i)} \sim q_\phi(z|x^{(i)})$.
2. Train a second κ-simple VAE, with independent parameters $\{\theta', \phi'\}$ and latent representation u, to learn the unknown distribution $q_\phi(z)$ as a new ground-truth distribution and use samples $\{z^{(i)}\}_{i=1}^n$ to learn it.
3. Samples approximating the original ground-truth μ_{gt} can then be formed via the extended ancestral process $u \sim \mathcal{N}(u|0, \mathrm{I}), z \sim p_{\theta'}(z|u), x \sim p_\theta(x|z)$

Two-Stage VAE Model

- If the first stage was successful, then even though they will not generally resemble $\mathcal{N}(z|0, \mathrm{I})$, samples from $q_\phi(z)$ will have nonzero measure across the full ambient space $\mathbb{R}^\kappa$.
- If $\kappa > r$, then the extra latent dimensions will be naturally filled in via randomness.
- Consequently, as long as we set $\kappa \geq r$, the operational regime of the second-stage VAE is effectively equivalent to the situation that the manifold dimension is equal to the ambient dimension.

Two-Stage VAE Model

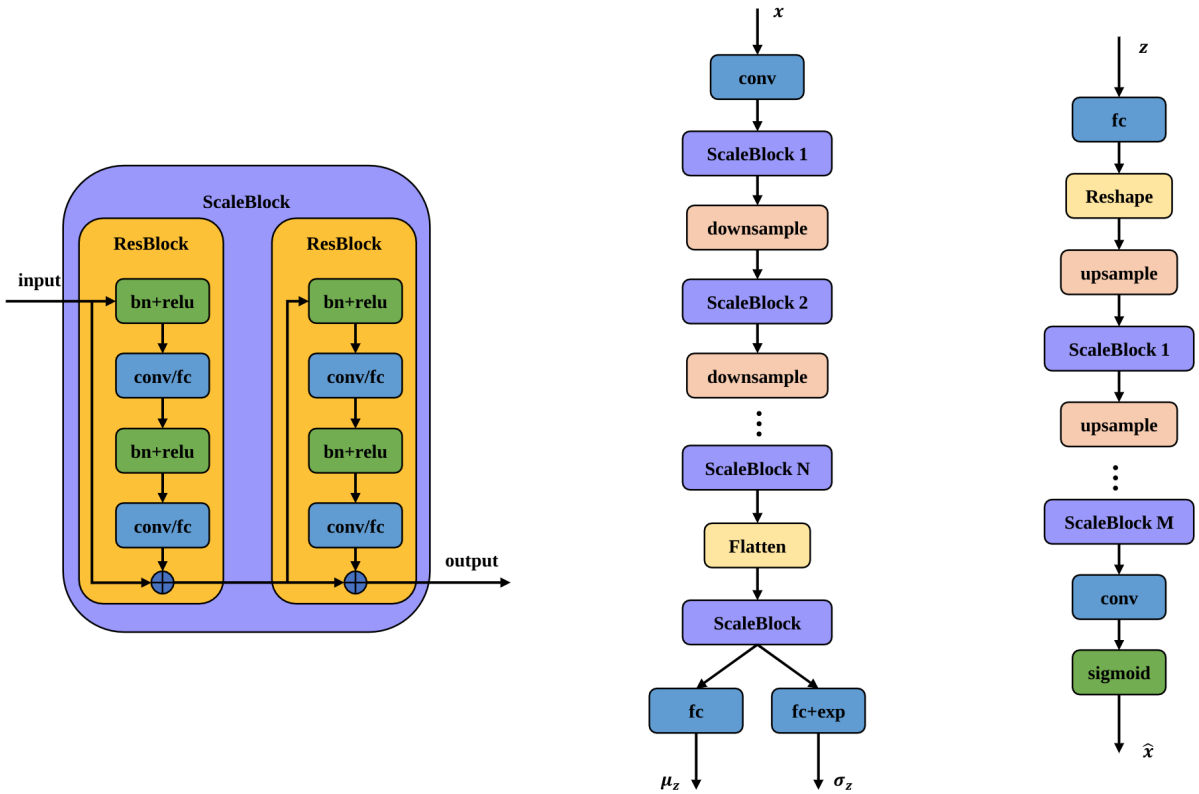


Figure: The structure of the first-stage of the Two-Stage VAE Model

Experiment Results

Evaluation Metrics:

Fréchet Inception Distance (FID) Score: used to assess the quality of images created by a generative model, comparing the generated images with the distribution of real images.

		MNIST	Fashion	CIFAR-10	CelebA
optimized, data-dependent settings	MM GAN	9.8 ± 0.9	29.6 ± 1.6	72.7 ± 3.6	65.6 ± 4.2
	NS GAN	6.8 ± 0.5	26.5 ± 1.6	58.5 ± 1.9	55.0 ± 3.3
	LSGAN	7.8 ± 0.6	30.7 ± 2.2	87.1 ± 47.5	53.9 ± 2.8
	WGAN	6.7 ± 0.4	21.5 ± 1.6	55.2 ± 2.3	41.3 ± 2.0
	WGAN GP	20.3 ± 5.0	24.5 ± 2.1	55.8 ± 0.9	30.3 ± 1.0
	DRAGAN	7.6 ± 0.4	27.7 ± 1.2	69.8 ± 2.0	42.3 ± 3.0
	BEGAN	13.1 ± 1.0	22.9 ± 0.9	71.4 ± 1.6	38.9 ± 0.9
default settings	Best default GAN	~ 10	~ 32	~ 70	~ 65
	VAE (cross-entr.)	16.6 ± 0.4	43.6 ± 0.7	106.0 ± 1.0	53.3 ± 0.6
	VAE (fixed γ)	52.0 ± 0.6	84.6 ± 0.9	160.5 ± 1.1	55.9 ± 0.6
	VAE (learned γ)	54.5 ± 1.0	60.0 ± 1.1	76.7 ± 0.8	60.5 ± 0.6
	VAE + Flow	54.8 ± 2.8	62.1 ± 1.6	81.2 ± 2.0	65.7 ± 2.8
	WAE-MMD	115.0 ± 1.1	101.7 ± 0.8	80.9 ± 0.4	62.9 ± 0.8
	2-Stage VAE (ours)	12.6 ± 1.5	29.3 ± 1.0	72.9 ± 0.9	44.4 ± 0.7

Note: The training of two stages need to be separate. Concatenating two stages and jointly training does not improve the performance.

Experiment Results

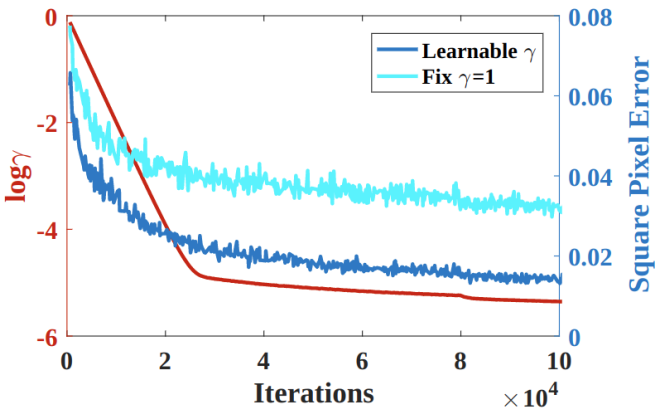
Evaluation Metrics:

Kernel Inception Distance (KID) applies a polynomial-kernel Maximum Mean Discrepancy (MMD) measure to estimate the inception distance, as FID score is believed to exhibit bias in certain circumstances.

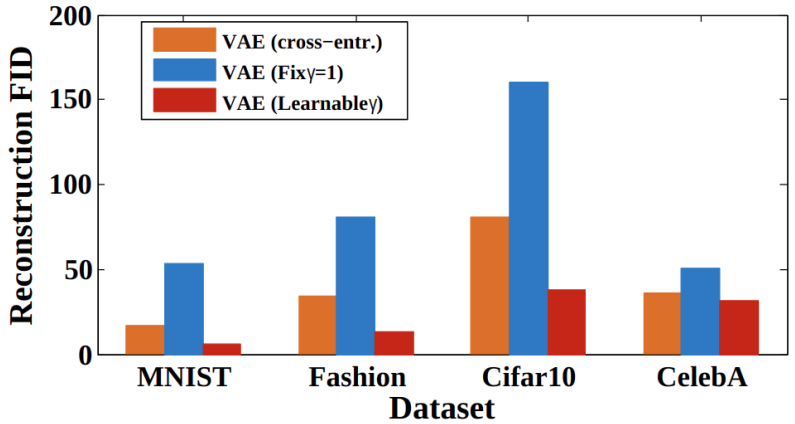
	MNIST	Fashion	CIFAR-10	CelebA
VAE (cross-entr.)	10.5 ± 0.3	37.0 ± 0.9	89.1 ± 1.3	48.4 ± 0.4
VAE (fixed γ)	42.7 ± 0.8	54.9 ± 0.8	153.4 ± 1.8	53.9 ± 0.7
VAE (learned γ)	51.5 ± 1.2	62.8 ± 1.7	64.6 ± 0.5	63.6 ± 1.1
VAE + Flow	56.0 ± 3.8	66.9 ± 1.6	68.5 ± 3.0	67.2 ± 3.4
WAE-MMD	137.8 ± 1.7	107.3 ± 1.5	58.7 ± 0.5	59.7 ± 0.8
2-Stage VAE (ours)	6.7 ± 0.3	25.9 ± 1.6	59.3 ± 0.9	40.9 ± 0.5

Analysis:

- The second stage of Two-Stage VAE model can reduce the gap between $q(z)$ and $p(z)$.
- γ will converge to zero at any global minimum of the VAE objective, allowing for tighter image reconstructions with better manifold fit.



(a) $\log\gamma$ and Reconstruction Error.



(b) Reconstruction FID.

Conclusion

Contribution of this paper

- Rigorously proved that VAE global optimum can in fact uniquely learn a mapping to the correct ground-truth manifold when $r < d$, but not necessarily the correct probability measure within this manifold.
- The proposed Two-Stage VAE model can resolve this issue and better recover the ground-truth manifold and reduce the gap between $p_\theta(z|x)$ and $q_\phi(z|x)$.
- The two-stage mechanism can improve the reconstruction of original distribution so that it has comparable performance with GAN models. This work narrows the gap between VAE and GAN models in terms of the realism of generated samples so that VAEs are worth considering in a broader range of applications.
- No need Gaussian assumption in the canonical VAE model to achieve the optimal solutions.

References

1. Dai, B., Wipf, D. (2019). Diagnosing and enhancing VAE models. arXiv preprint arXiv:1903.05789.
2. Carl Doersch. Tutorial on variational autoencoders.arXiv:1606.05908, 2016
3. Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and SeppHochreiter. GANs trained by a two time-scale update rule converge to a local Nashequilibrium. InAdvances in Neural Information Processing Systems, pages 6626–6637,2017.
4. Miko laj Bi nkowski, Dougal J Sutherland, Michael Arbel, and Arthur Gretton. Demystifyingmmd gans.arXiv:1801.01401, 2018