AS-IntroVAE: Adversarial Similarity Distance Makes Robust IntroVAE

Changjie Lu[1], *ShenZheng*[1,3], *ZiruiWang*[2], *OmarDib*[1], *GauravGupta*[1]

Wenzhou Kean University, Zhejiang University, Carnegie Mellon University

Introspective VAE(Intro-VAE)[Hua+18]

Combine VAE(statistical analysis) and GAN(adversarial learning) together.

$$\begin{aligned} \mathcal{L}_E &= ELBO(x) + \sum_{s=r,g} \left[m - KL\left(q_\phi\left(z|x_s\right) \| p(z)\right]^+ \right. \\ \mathcal{L}_D &= \sum_{s=r,g} \left[KL\left(q_\phi\left(z|x_s\right) \| p(z)\right)\right] \end{aligned} \tag{1}$$

where x_r is the reconstructed image, x_g is the generated image, and m is the hard threshold for constraining the KL divergence.

Soft-IntroVAE[DT21]

The hard threshold makes training stability sensitive to the hyper parameter, S-IntroVAE introduces a soft expression.

$$\begin{aligned}\mathcal{L}_E &= ELBO(x) - \frac{1}{\alpha}\sum_{s=r,g}\exp\left(\alpha ELBO\left(x_s\right)\right) \\ \mathcal{L}_D &= ELBO(x) + \gamma\sum_{s=r,g} ELBO\left(x_s\right)\end{aligned} \tag{2}$$

where α, γ are both hyperparameters.

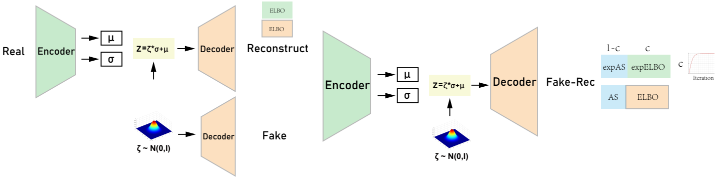

Figure 1: AS-IntroVAE workflow. In the first phase, the encoder-decoder receives the real image and produce the reconstructed image. In the second phase, the **same** encoder-decoder conduct adversarial learning in the latent space for the reconstructed image and the fake image.

Limitation and Solution

Those introspective learning-based methods suffer from the **posterior collapse** problem and the **vanishing gradient** problem.

Contribution:

1. A new introspective variational autoencoder named Adversarial Similarity Distance Introspective Variational Autoencoder (AS-IntroVAE)
2. A new theoretical understanding of the posteriors collapse and the vanishing gradient problem in VAEs.
3. A novel similarity distance named Adversarial Similarity Distance (AS-Distance) for measuring the differences between the real and the synthesized images.
4. Promising results on image generation and image reconstruction tasks with significantly faster convergence speed

Inspired by 1-Wasserstein distance, which could provide stable gradients, the AS-Distance is defined as:

$$D\left(p_r, p_g\right) = \mathbb{E}_{x \sim p(z)}\left[\left(\mathbb{E}_{x \sim p_r}\left[q\left(z|x\right)\right] - \mathbb{E}_{x \sim p_g}\left[q\left(z|x\right)\right]\right)\right]^2 \quad (3)$$

where p_r is distribution of real data, p_g is distribution of generated data. The encoder and the decoder plays an adversarial game on this distance:

$$arg \min_{Dec} \max_{Enc} D\left(p_r, p_g\right) \quad (4)$$

We use 2-Wasserstein so that we could apply a kernel trick on Equ.3.

$$D(p_r, p_g) = \mathbb{E}_{x\sim p_{r,g}}\left[k\left(x_r^i, x_r^j\right) + k\left(x_g^i, x_g^j\right) - 2k\left(x_r^i, x_g^j\right)\right] \tag{5}$$

where $k\left(x_r^i, x_g^j\right) = \mathbb{E}_{z\sim p(z)}[q(z|x_r^i)\cdot q(z|x_g^i)]$.
Since the latent space is a normal distribution. This kernel k can be deduced as

$$k\left(x_r^i, x_g^j\right) = \frac{-\frac{1}{2}\frac{\left(u_r^i - u_g^j\right)^2}{\lambda_r^i + \lambda_g^j}}{(2\pi)^{\frac{n}{2}}\cdot\left(\lambda_r^i + \lambda_g^j\right)^{\frac{1}{2}}} \tag{6}$$

where u, λ represent the variational inference on the mean and variance of x, i, j represent the ith, jth pixel in images.

Inspired by ([Fu+19]), we decide to gradually increase the weight for KL (from 0 to 1), and decrease the weight for AS (from 1 to 0) during training.
We derive the loss function for AS-IntroVAE as:

$$
\begin{aligned}
\mathcal{L}_{E_\phi} &= ELBO(x) - \frac{1}{\alpha} \sum_{s=r,g} \exp(\alpha(\mathbb{E}_{q(z|x_s)}[\log p(x \mid z)] \\
&\quad + cKL(q_\phi(z|x_s)\|p(z)) + (1-c)D(x_r, x_g))) \\
\mathcal{L}_{D_\theta} &= ELBO(x) + \gamma \sum_{s=r,g} (\mathbb{E}_{q(z|x_s)}[\log p(x \mid z)] \\
&\quad + cKL(q_\phi(z|x_s)\|p(z)) + (1-c)D(x_r, x_g))
\end{aligned}
\tag{7}
$$

where $c = min(i * 5/T, 1)$, i is the current iteration and T is total iteration.

Results

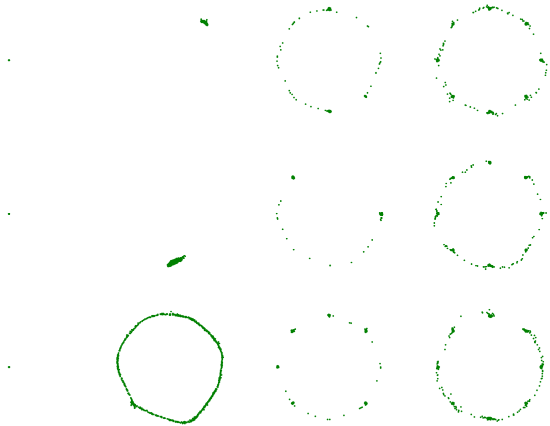

Figure 2: Visual Comparison on 2D Toy Dataset 8 Gaussians. From top to bottom row: results with different hyperparameters. From left to right column: VAE, IntroVAE, S-IntroVAE, Ours.

Figure 3: Image Generation Visual Comparison at CelebA-128 dataset. From left to Right: WGAN-GP, S-IntroVAE, Ours

Figure 4: Image Generation Visual Comparison at CelebA-256 dataset.

Figure 5: Image Reconstruction Visual Comparison at CelebA-128 dataset.

		VAE	IntroVAE	S-IntroVAE	Ours
2*C1	KL	220.2	192.4	50.2	**3.4**
	JSD	110.1	56.0	16.9	**5.6**
2*C2	KL	220.3	191.1	136.5	**1.3**
	JSD	110.0	68.0	36.6	**4.4**
2*C3	KL	220.2	64.0	46.2	**2.0**
	JSD	109.8	53.0	9.6	**7.1**

Table 1: 2D Toy Dataset 8 Gaussians Score KL↓/JSD↓ Table

	WGAN-GP	S-IntroVAE	Ours
MNIST	139.02	98.84	**96.16**
CIFAR-10	434.11	275.20	**271.69**
CelebA-128	160.53	140.35	**130.74**
CelebA-256	170.79	143.33	**129.61**

Table 2: Image Generation FID Score↓ Table.

	PSNR		SSIM		MSE	
	S-IntroVAE	Ours	S-IntroVAE	Ours	S-IntroVAE	Ours
MNIST	20.282	**21.014**	0.885	**0.898**	0.011	**0.009**
CIFAR-10	19.300	**19.445**	0.599	**0.620**	**0.019**	**0.019**
Oxford	15.372	**20.168**	0.348	**0.604**	0.049	**0.013**
CelebA-128	17.818	**22.924**	0.561	**0.801**	0.018	**0.006**
CelebA-256	22.422	**23.156**	**0.790**	0.758	0.007	**0.006**

Table 3: Image Reconstruction PSNR↑/SSIM↑/MSE↓ Score Table

Figure 6: The training stability visual comparison at CelebA-128 dataset. From left to right panel: 10 epoch, 20 epoch, 50 epoch.

Figure 7: Image generation visual comparisons at CelebA-128 dataset (resolution: 128 $\times$ 128).

Figure 8: Image generation visual comparisons at CelebA-256 dataset (resolution: 256×256).

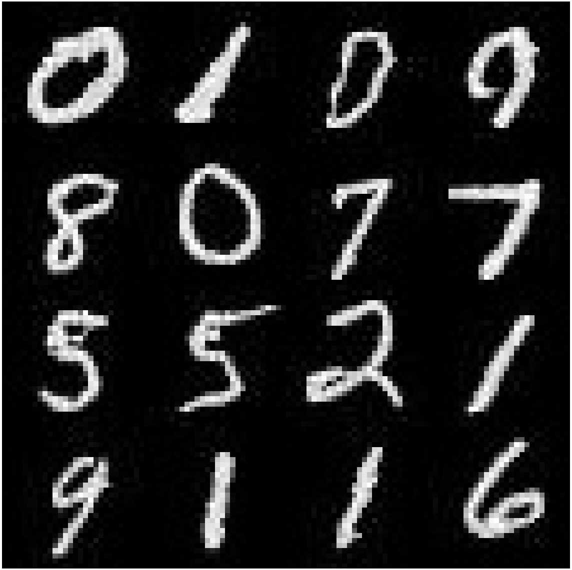

Figure 9: Image generation visual comparisons at MNIST dataset (resolution: 28×28).