AVAE: Adversarial Variational Auto Encoder

Antoine Plumerault (CEA/Centrale-Supelec), Hervé Le Borgne (CEA), Céline Hudelot (Centrale-Supelec)

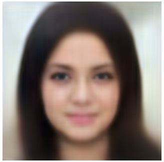
December 9, 2020

Introduction

Two main types of generative models

- VAEs have several advantages over GANs

GAN	VAE
+ realistic images	+ disentangled latent space
− mode collapse	+ encoder model
− difficult to invert	+ easy to train
	− blurry images



Problematic: VAE fail to produce realistic images (w.r.t GANs)

- How can we explain this lack of realism ?
- Can we combine the best of VAEs and GANs ?

Understanding VAEs and GANs

Which problem for VAEs to produce realistic images ?

1. Information bottleneck:

$$\mathcal{L}_{\text{VAE}} = \underbrace{\mathbb{E}\left[\mathbb{E}_{q_{\theta_e}(z|x)}\left[-\log p_{\theta_d}(x|z)\right]\right]}_{\text{reconstruction error}} + \underbrace{I_\theta(x;z)}_{\text{mutual information}} + \underbrace{\text{KL}(p_{\theta_e}(z)||p(z))}_{\text{prior on } z} \tag{1}$$

   $\rightarrow$ incomplete information
   $\rightarrow$ mean value of all possible images
   $\rightarrow$ blurry results

2. Underestimation of natural image manifold dimensionality:

   $\rightarrow$ approximation of the manifold with a simpler one
   $\rightarrow$ uncertainty on other dimensions responsible of smaller variations (e.g. textures)
   $\rightarrow$ mean value of all possible images
   $\rightarrow$ blurry results (no texture in images)

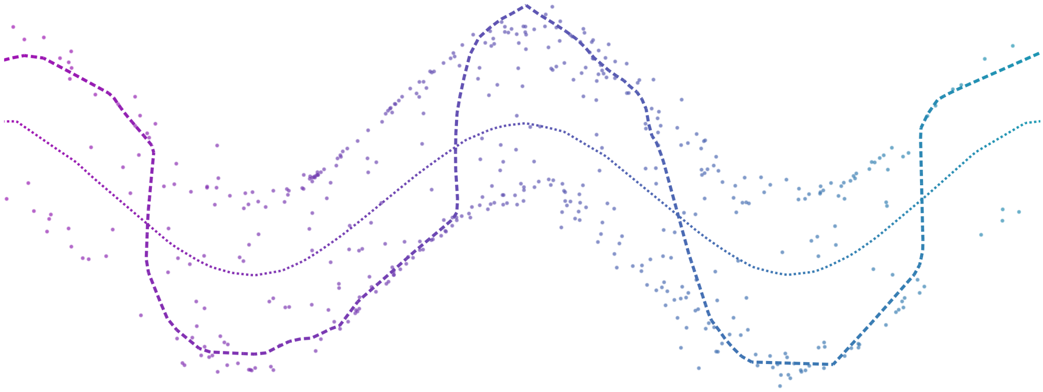
How GANs are able to produce realistic images

GANs also underestimate the dimension of the natural image manifold.

$\rightarrow$ Question: How are they able to produce realistic images ?

$\rightarrow$ Answer: Mode collapse ! $\rightarrow$ only a few but plausible texture configurations are generated.

Illustration on a toy example

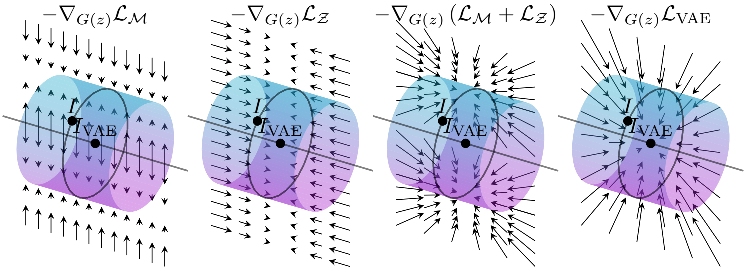


dots: data points dotted line: VAE manifold dashed line: GAN manifold

How to solve the VAE problem ?

Objective: Create a reconstruction error $\mathcal{L}_{\mathcal{Z}}$:

- that is powerful enough to favor accurate reconstructions.
- that does not favor blurry reconstruction to allow realistic reconstructions.



- cylinders: real data high-dimensional manifold
- black line: low-dimensional manifold of VAEs reconstructions
- arrows: gradient of different losses

What properties such a reconstruction loss should satisfy ?

With reconstruction errors of the form $\mathcal{L}_{\mathcal{Z}}(\hat{x}, x) = \frac{1}{2}||f(\hat{x}) - g(x)||^2$ where:

- f is an arbitrary differentiable function
- g is a stochastic function

Optimal solutions $\hat{x}^*(z)$ verifies:

$$\boxed{f(\hat{x}^*(z)) = \mathbb{E}_{g(x)\sim p_{\theta_e}(g(x)|z)}\left[g(x)\right]} \tag{2}$$

- $f(\hat{x})$ should carry the maximum of information about $\hat{x}$ and $g(x)$ should be close to $f(x)$.
- Common optimum with the GAN objective $\iff$ $p(f(\hat{x}^*(z))) = p(f(x))$ for $z \sim p(z)$ and $x \sim p_{\mathcal{D}}(x)$.

A simple example: the MSE

$\mathcal{L}_{\mathcal{Z}}(\hat{x}, x) = MSE(\hat{x}, x) = \frac{1}{2}||\hat{x} - x||^2 \rightarrow$ optimal solution:

$$\boxed{\hat{x}^*(z) = \mathbb{E}_{x \sim p_{\theta_e}(x|z)}[x]} \tag{3}$$

- $f(\hat{x})$ carry all the information about $\hat{x}$ as it is the identity, and $g(x) = f(x)$.
- Optimal solution = mix of likely solutions $\rightarrow$ blurry / unrealistic image.
  $p(f(\hat{x}^*(z))) = p(\hat{x}^*(z)) \neq p(x) = p(f(x))$ for $z \sim p(z)$ and $x \sim p_{\mathcal{D}}(x)$.

The AVAE framework

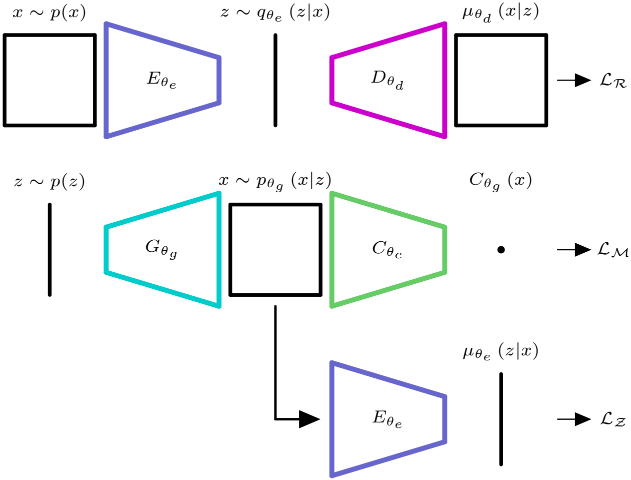
Our loss

With: $f(\hat{x}) = \frac{\mu_{\theta_e}(\hat{x})}{\sigma_{\theta_e}}$, $g(x) = \frac{\sqrt{1-\sigma_{\theta_e}^2}}{\sigma_{\theta_e}} z$ $\rightarrow$

$$\boxed{\mathcal{L}_{\mathcal{Z}}(\hat{x}) = \frac{1}{2} \left\| \frac{\mu_{\theta_e}(x) - \sqrt{1 - \sigma_{\theta_e}^2} z}{\sigma_{\theta_e}} \right\|^2} \quad (4)$$

- $f(\hat{x})$ carry the information about $\hat{x}$ contained in z , and $g(x) = \frac{\sqrt{1-\sigma_{\theta_e}^2}}{\sigma_{\theta_e}} z \approx \frac{\mu_{\theta_e}(x)}{\sigma_{\theta_e}} = f(x)$
- $\mu_{\theta_e}(\hat{x}^*(z)) = \sqrt{1 - \sigma_{\theta_e}^2} z \rightarrow p(\mu_{\theta_e}(\hat{x}^*(z))) = \mathcal{N}(\mu_{\theta_e}(x); 0, I - \Sigma) = p(\mu_{\theta_e}(x))$.

Full AVAE framework



Results

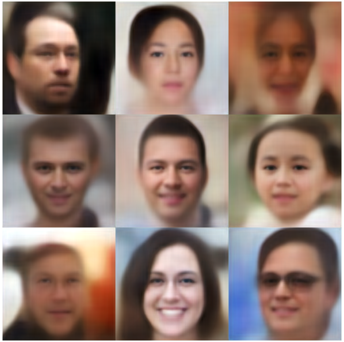
Quantitative results on CelebA

metric	VAE	GAN	VAE/GAN	BiGAN	Ours
mse $\downarrow$	0.03 ± 0.00	–	0.07 ± 0.00	0.18 ± 0.01	0.05 ± 0.00
lpips $\downarrow$	0.18 ± 0.00	–	0.09 ± 0.00	0.16 ± 0.00	0.11 ± 0.00
fid $\downarrow$	60.04 ± 0.47	14.54 ± 0.41	26.45 ± 4.66	18.49 ± 5.06	15.01 ± 0.82

- MSE: favorable to VAE a priori.
- MSE: favorable to our approach a priori.
- LPIPS & FID: favorable to VAE/GAN a priori.

Qualitative results

original images



VAE decoder reconstructions

generator reconstructions

Conclusion