Simple and Effective VAE Training with Calibrated Decoders

2022/09/29

Ping-Han Hsieh

ELBO of Variational Autoencoder

Gaussian PDF

$$\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

ELBO

$$\ln p_\theta(x) \geq \mathbb{E}_{q_\phi(z|x)}\left[\ln p_\theta(x|z)\right] + \mathrm{D}_{\mathrm{KL}}(q_\phi(z|x)||p_\theta(z))$$

Gaussian data distribution (w/ unit variance) $p_\theta(x|z) = \mathcal{N}(\mu_\theta(z), I)$

$$-\ln p(x|z) = \frac{1}{2}||\hat{x}-x||^2 + D\ln\sqrt{2\pi} = \frac{1}{2}||\hat{x}-x||^2 + c \;= \frac{D}{2}\mathrm{MSE}(\hat{x},x) + c$$

Gaussian data distribution (w/ full diagonal variance) $p_\theta(x|z) = \mathcal{N}\left(\mu_\theta(z), \sigma_\theta(z)^2\right)$

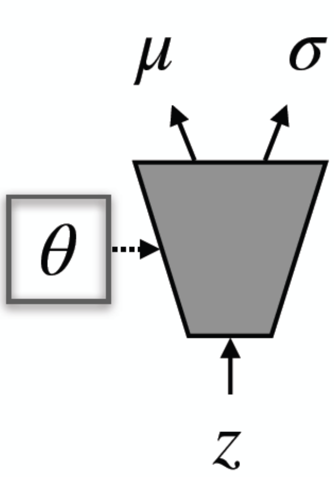
$$-\ln p(x|z) = \frac{1}{2\sigma^2}||\hat{x}-x||^2 + D\ln\sigma\sqrt{2\pi} = \frac{1}{2\sigma^2}||\hat{x}-x||^2 + D\ln\sigma + c = D\ln\sigma + \frac{D}{2\sigma^2}\mathrm{MSE}(\hat{x},x) + c.$$

Gaussian data distribution (w/ shared variance) $p_{\theta,\sigma}(x|z) = \mathcal{N}\left(\mu_\theta(z), \sigma^2 I\right)$

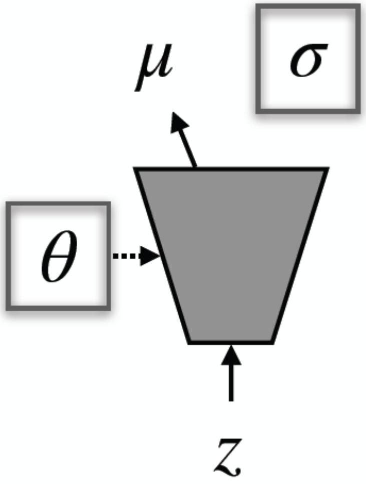
$$-\ln p(x|z) = \frac{1}{2\sigma^2}||\hat{x}-x||^2 + D\ln\sigma\sqrt{2\pi} = \frac{1}{2\sigma^2}||\hat{x}-x||^2 + D\ln\sigma + c = D\ln\sigma + \frac{D}{2\sigma^2}\mathrm{MSE}(\hat{x},x) + c.$$

* share σ across images

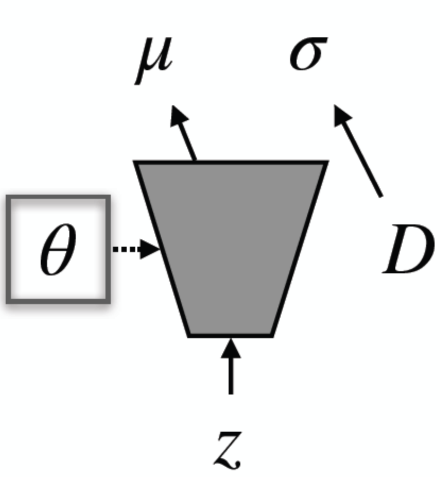
Decoder



σ-VAE
w/ full diagonal variance



σ-VAE
w/ shared variance



optimal σ-VAE
w/ shared variance
(analytical solution)

Connection to β-VAE

ELBO (β-VAE w/ unit variance)

$$\mathcal{L}^{\beta} = \frac{D}{2} MSE(\hat{x}, x) + \beta D_{KL}(q(z|x)||p(z))$$

ELBO (σ-VAE)

$$\mathcal{L}_{\theta,\phi,\sigma} = D \ln \sigma + \frac{D}{2\sigma^2} MSE(\hat{x}, x) + D_{KL}(q(z|x)||p(z))$$

$$\text{let } \sigma^2 = \beta$$

$$\mathcal{L}_{\theta,\phi,\beta} = \frac{D}{2\beta} MSE(\hat{x}, x) + D_{KL}(q(z|x)||p(z)) + c_1$$

$$\beta\mathcal{L}_{\theta,\phi,\beta} = \frac{D}{2} MSE(\hat{x}, x) + \beta D_{KL}(q(z|x)||p(z)) + c_2$$

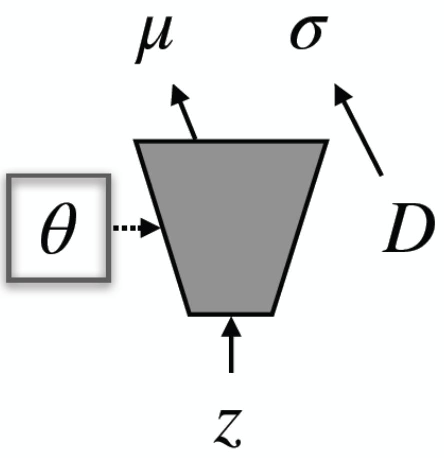
Optimal σ-VAE

- The maximum likelihood estimation of variance:

$$\sigma^{*2} = \arg\max_{\sigma^2} \mathcal{N}(x|\mu, \sigma^2 I) = \mathrm{MSE}(x, \mu)$$
$$\mathrm{MSE}(x, \mu) = \frac{1}{D}\sum_i (x_i - \mu_i)^2$$

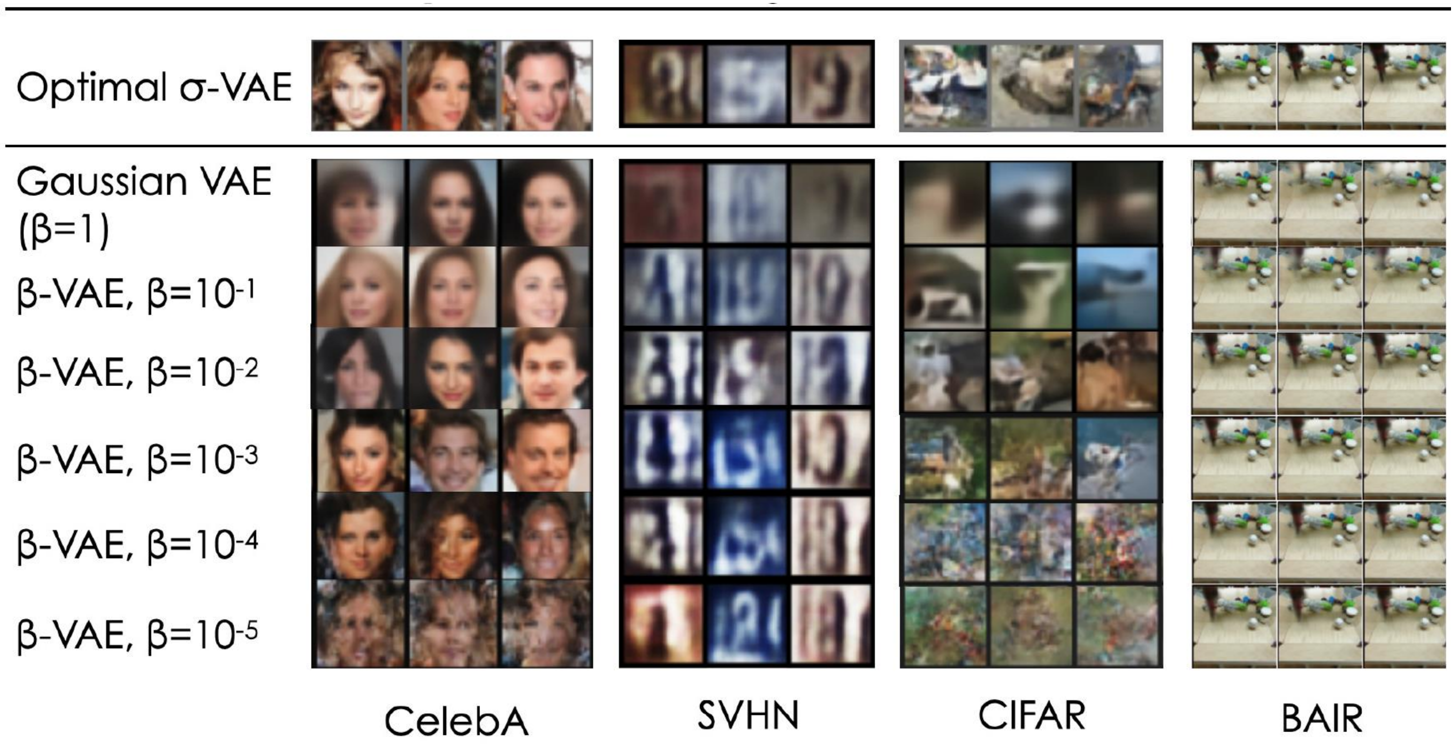
- Use batchwise estimation during mini-batch training.
- Using the running average of the variance over training during testing.

$$\sigma^* = \arg\max_{\sigma} \mathbb{E}_{x\sim\mathrm{Data}}\mathbb{E}_{q(z|x)}\left[\ln p(x|\mu_\theta(z), \sigma^2 I)\right]$$
$$= \mathbb{E}_{x\sim\mathrm{Data}}\mathbb{E}_{q(z|x)}\mathrm{MSE}(x, \mu_\theta(z)).$$



optimal σ-VAE
w/ shared variance
(analytical solution)

Compare with β-VAE



	β	$-\log p \downarrow$	FID $\downarrow$
β-VAE	0.001	< 21.43	44.54
β-VAE	0.01	< -3186	27.93
β-VAE	0.1	< -1223	28.3
β-VAE	1	< 1381	70.39
β-VAE	10	< 4056	219.3
σ-VAE	0.006	$< \mathbf{-3333}$	**22.25**

$$d_F(\mu, \nu) := \left(\inf_{\gamma \in \Gamma(\mu,\nu)} \int_{\mathbb{R}^n \times \mathbb{R}^n} \|x - y\|^2 \, \mathrm{d}\gamma(x, y) \right)^{1/2}$$

Impact on Latent Variables

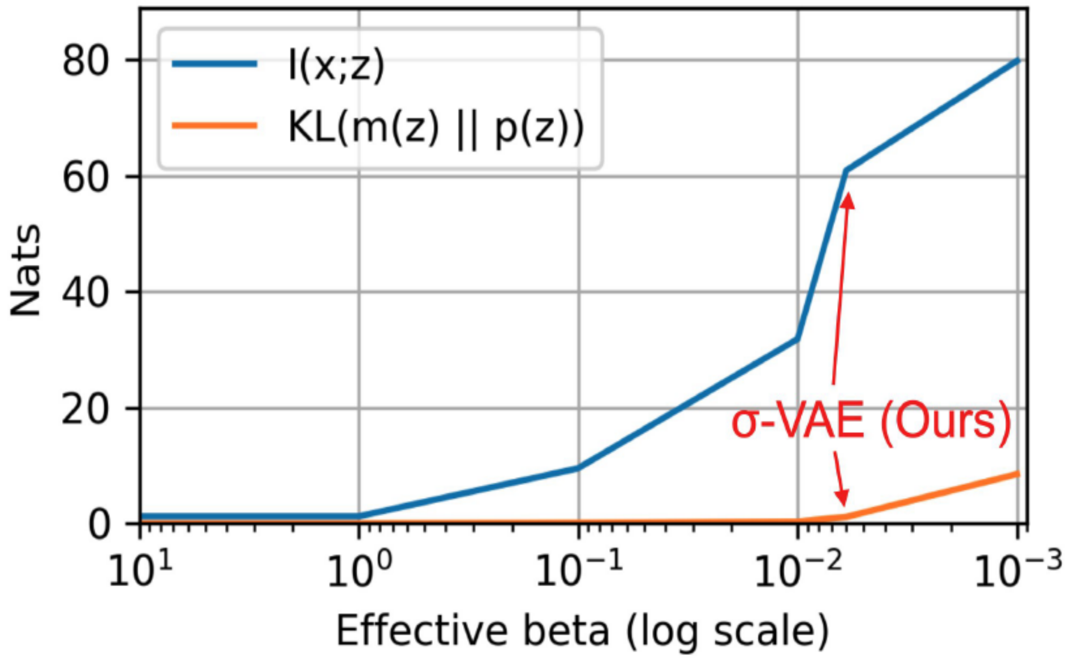
- Decomposition of the KL-Divergence

$$
\begin{aligned}
&E_{p_d(x)}\left[D_{KL}(q(z|x)||p(z))\right] \\
&= E_{p_d(x)}\left[D_{KL}(q(z|x)||m(z))\right] + D_{KL}(m(z)||p(z)) \\
&= I_e(x; z) + D_{KL}(m(z)||p(z)).
\end{aligned}
$$

where $m(z) = E_{p_d(x)} q(z|x)$

	β	$-\log p \downarrow$	FID $\downarrow$
β-VAE	0.001	< 21.43	44.54
β-VAE	0.01	< -3186	27.93
β-VAE	0.1	< -1223	28.3
β-VAE	1	< 1381	70.39
β-VAE	10	< 4056	219.3
σ-VAE	0.006	$< \mathbf{-3333}$	**22.25**

σ-VAE captures the inflection point



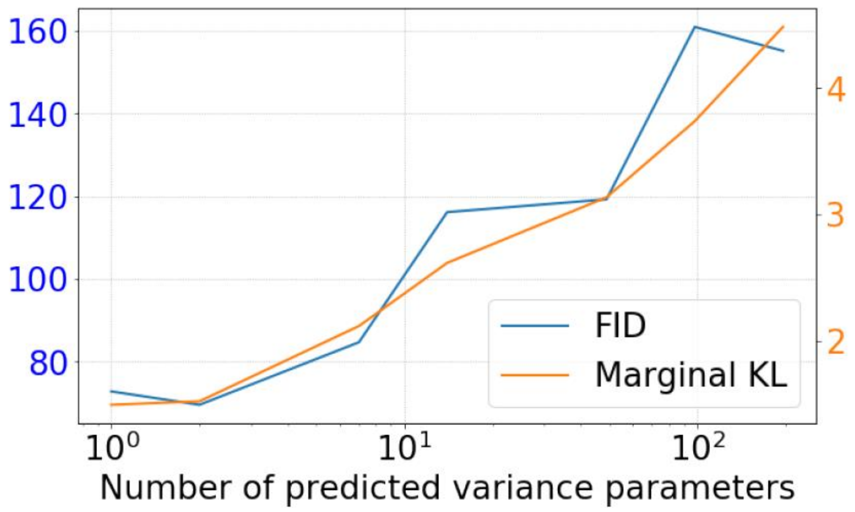
Compare with Other (Calibration) Methods

	CelebA HVAE		SVHN VAE		CIFAR HVAE		BAIR SVG	
	$-\log p \downarrow$	FID $\downarrow$	$-\log p \downarrow$	FID $\downarrow$	$-\log p \downarrow$	FID $\downarrow$	$-\log p \downarrow$	FID $\downarrow$
Bernoulli VAE [1]		177.6		43.26		284.5		122.6
Categorical VAE	$< \mathbf{6359}$	71.5	< 9179	46.13	$< \mathbf{7179}$	**101.7**	N/A	N/A
Bitwise-categorical VAE	< 9067	66.61	< 10800	33.84	< 9390	**91.2**	< 48744	46.13
Logistic mixture VAE	< 7932	65.3	$< \mathbf{9085}$	43.19	< 8443	143.1	$< \mathbf{40616}$	42.94
Gaussian VAE	< 7173	186.5	< 2184	112.5	< 7186	293.7	< -10379	35.64
Per-pixel σ-VAE	< -7814	159.3	< -3592	114.7	< -7222	131	< -14051	41.98
Student-t VAE [2]	< -8401	71.06	$< \mathbf{-3659}$	70.4	$< \mathbf{-7419}$	123.6	-	-
β-VAE [3]	< -2713	**61.6**	< -3186	27.93	< -331	**103**	< -13472	34.64
Shared σ-VAE	< -6374	**60.7**	< -3349	**22.25**	< -5435	116.1	< -13974	34.24
Optimal σ-VAE	$< \mathbf{-8446}$	**60.3**	< -3333	27.25	< -5677	**101.4**	$< \mathbf{-14173}$	34.13
Opt. per-image σ-VAE		66.01		26.28		**104.0**		**33.21**

Common Challenges on Variance Calibration

- Numerical instability (extremely low variance)
- Bounding variance leads to poor generative results despite high ELBO.
- Shared variance, per-image variance, and per-pixel variance lead to different performance.
  - More expressive variance decoder produces unrealistic samples and meaningless latent representations.
  - Sharing variance across images is a useful inductive bias.

per-pixel σ-VAE
w/ partially shared variance



Optimal σ-VAE Improves Learning Process

- Speed up learning process.
- Per-image optimal σ-VAE achieves best image quality.
- Analytical solution for the optimal σ-VAE makes it easy to implement per-image, per-pixel shared variance decoders.