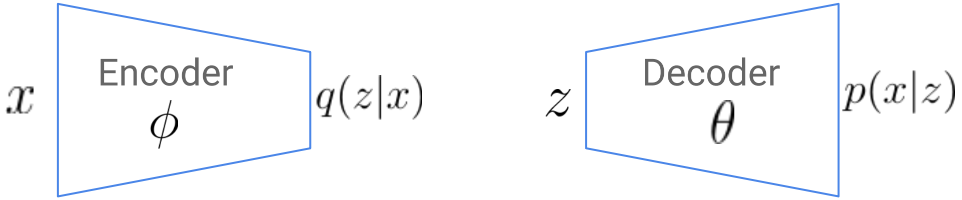
On Implicit Regularization in β-VAEs

Abhishek Kumar, Ben Poole

Google Research, Brain Team

ICML 2020

β-VAE



$$\max_{q \in \mathcal{Q}, p_{x|z}} \mathbb{E}_x[\mathbb{E}_{q(z|x)} \log p_{x|z}(x|z) - \beta \, KL(q(z|x) \| p_z(z))]$$

How does variational family $\mathcal{Q}$ regularize the learned generative model?

- Uniqueness of learned generative model (global regularization)
- Influencing the local geometry of the decoding model $p(x|z)$
- Deterministic approximation of β-VAE
- Empirical validation of theory and accuracy of approximations

Latent variable models and Uniqueness

Fixed prior $p(z)$, conditional decoding model $p(x|z)$, marginal $p(x) = \int dz\, p(z)p(x|z)$

$$z \xrightarrow{\quad r \quad} z' \quad \text{such that} \quad p(z) = p(z')$$

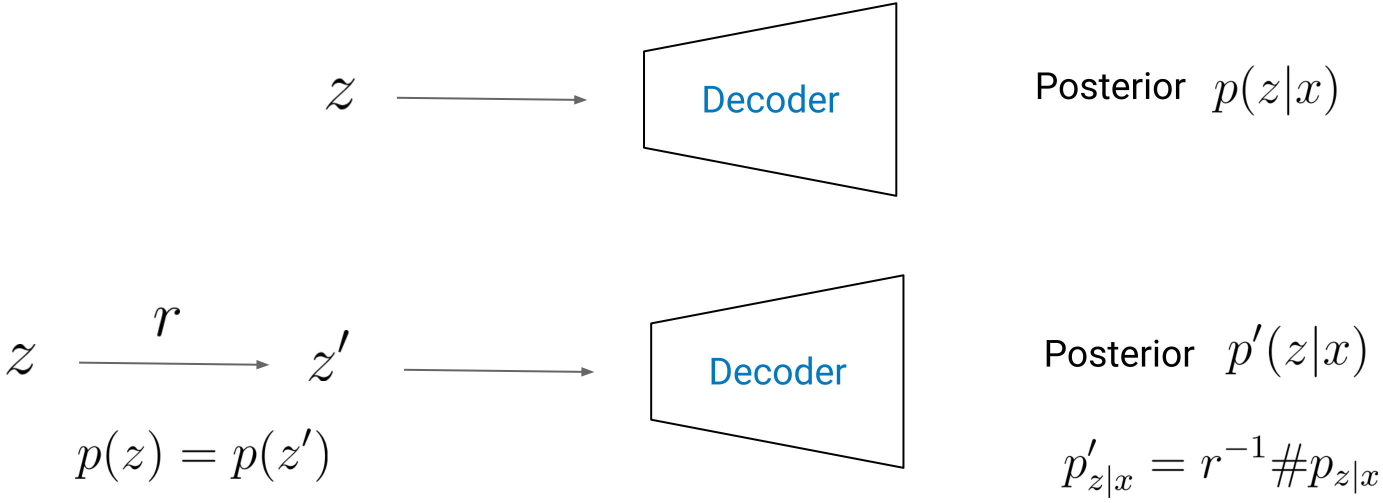
A set of solutions (latent representations) that are equivalent in terms of marginal likelihood [1].

Uniqueness: Ignoring permutations and transformations that act separately on each latent

[1] Locatello et al, Challenging Common Assumptions in the Unsupervised Learning of Disentangled Representations, 2019.

Uniqueness via variational family

When true posterior $p(z|x)$ is in the variational family $\mathcal{Q}$:



If $p_{z|x} \in \mathcal{Q}$ but $p'_{z|x} \notin \mathcal{Q}$, maximum ELBO for the transformed model will be less than the untransformed model.

Example: Isotropic Gaussian prior and orthogonal transforms ($p_{z|x} \in \mathcal{Q}$)

Isotropic Gaussian $p(z)$ is invariant under orthogonal transformations $r \in R$

- Transforming latents by orthogonal transforms will leave the marginal $p(x)$ unchanged

Restricting variational family to mean-field can break this orthogonal “symmetry”:

- If $p_{z|x}$ is mean-field, $r\#p_{z|x}$ will be mean-field, if and only if (Darmois, 1953; Skitovitch, 1953)

  (i) $p_{z|x}$ is factorized Gaussian, and

  (ii) variances of $p_{z|x}$ are all equal (isotropic)

Models with non-Gaussian factorized $p_{z|x}$ will not have non-uniqueness to orthogonal transforms.

Uniqueness via variational family (when $p_{z|x} \notin \mathcal{Q}$)

Choice of $\mathcal{Q}$ can lead to uniqueness even if $p_{z|x} \notin \mathcal{Q}$.

R: set of transforms w.r.t. which we want uniqueness (that leave the prior invariant)

$\widetilde{\mathcal{Q}}$: completion of $\mathcal{Q}$ by R := $\{r\#q : q \in \mathcal{Q}, r \in R\} \cup \mathcal{Q}$

Two conditions:

1. If $\mathcal{Q}$ is such that $q \in \mathcal{Q} \Rightarrow r\#q \notin \mathcal{Q}$, and
2. $\arg\max_{q \in \widetilde{\mathcal{Q}}} \mathrm{ELBO}(q, p^*)$ is unique (holds when $\widetilde{\mathcal{Q}}$ is convex)

Then transforming the latents with $r \in R$ will result in reducing the β-VAE objective value.

Uniqueness via variational family (when $p_{z|x} \notin \mathcal{Q}$)

Generative model of data: $x = \nu(Wz) + \epsilon, \quad z \sim \mathcal{N}(0, I), \epsilon \sim \mathcal{N}(0, .05^2)$

Train a VAE on this data:

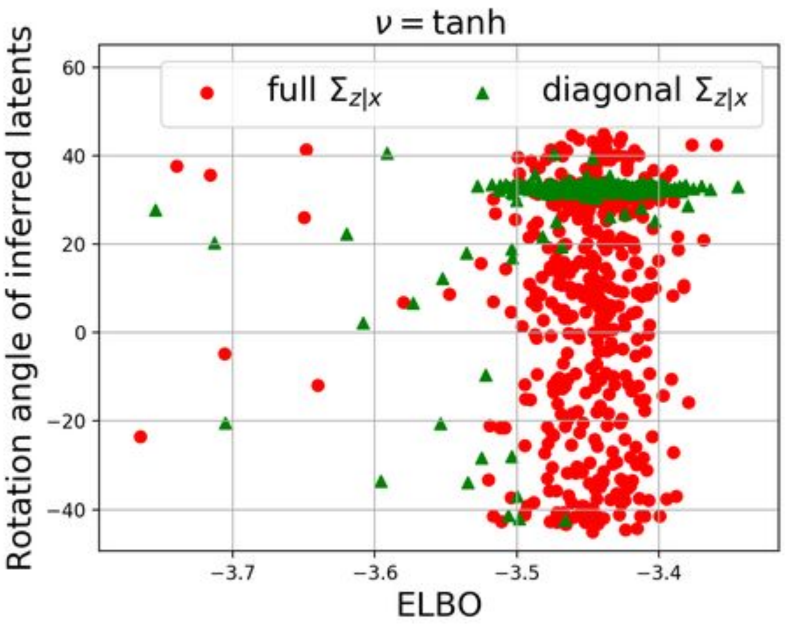
- Decoder $x = \nu(Az) + b$
- Encoder $q(z|x) = \mathcal{N}(Cx + d, \Sigma)$

Uniqueness via variational family (when $p_{z|x} \notin \mathcal{Q}$)

Generative model of data: $x = \nu(Wz) + \epsilon, \quad z \sim \mathcal{N}(0, I), \epsilon \sim \mathcal{N}(0, .05^2)$

Train a VAE on this data:

- Decoder $x = \nu(Az) + b$
- Encoder $q(z|x) = \mathcal{N}(Cx + d, \Sigma)$

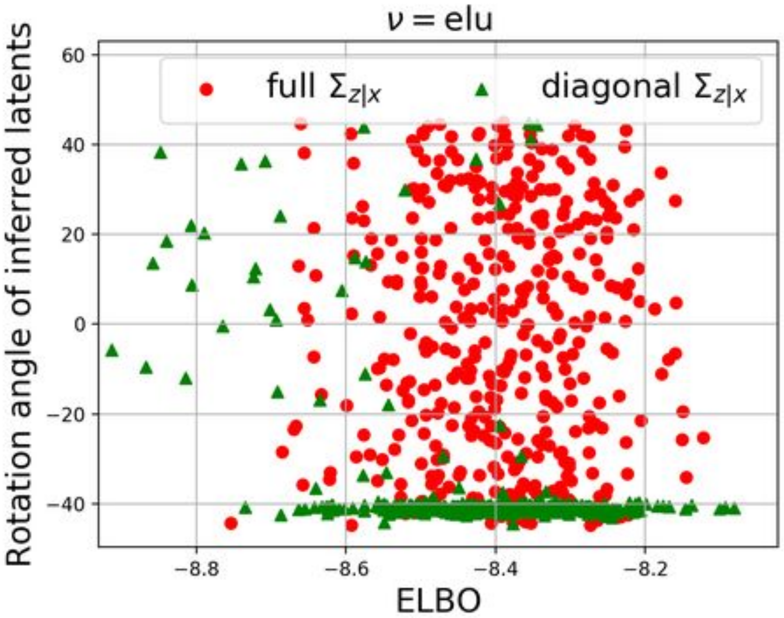


Uniqueness via variational family (when $p_{z|x} \notin \mathcal{Q}$)

Generative model of data: $x = \nu(Wz) + \epsilon, \quad z \sim \mathcal{N}(0, I), \epsilon \sim \mathcal{N}(0, .05^2)$

Train a VAE on this data:

- Decoder $x = \nu(Az) + b$
- Encoder $q(z|x) = \mathcal{N}(Cx + d, \Sigma)$



More details in the paper on implications for disentanglement [1].

[1] Locatello et al, Challenging Common Assumptions in the Unsupervised Learning of Disentangled Representations, 2019.

Regularization: local geometry

How does variational family regularize the local geometry of the generative model?c

Assumption: First two moments exist for $q(z|x)$.

Let

$$f_x(z) = \log p(x|z) \qquad \text{and} \qquad \mu_{z|x} = \mathbb{E}_{q_\phi(\cdot|x)}[z] = h_\phi(x)$$

Taylor approximation around $\mu_{z|x}$:

$$f_x(z) = \log p(x|z) \approx \log p(x|h_\phi(x)) + \underbrace{J_{f_x}}_{\text{Jacobian}}(h_\phi(x))(z - h_\phi(x)) + \frac{1}{2}(z - h_\phi(x))^\top \underbrace{H_{f_x}}_{\text{Hessian}}(h_\phi(x))(z - h_\phi(x)),$$

Regularization: local geometry

$$f_x(z) = \log p(x|z) \approx \log p(x|h_\phi(x)) + J_{f_x}(h_\phi(x))(z - h_\phi(x)) + \frac{1}{2}(z - h_\phi(x))^\top \underbrace{H_{f_x}(h_\phi(x))}_{\text{Hessian}}(z - h_\phi(x)),$$

Taking expectation wrt. $q(z|x)$, Taylor approximation of β-VAE reduces to

$$\log p_\theta(x|h_\phi(x)) + \frac{1}{2}tr(H_{f_x}(h_\phi(x))\underbrace{\Sigma_{z|x}}_{\text{Covariance of } q(z|x)}) - \beta\, KL(q_\phi(z|x) \| p(z))$$

Regularization: local geometry

We can further reduce the approximation in terms of the Jacobian of the decoder.

$$\underbrace{H_{f_x}(z)}_{:= \nabla_z^2 \log p(x|z)} \approx J_g(z)^\top \underbrace{H_{p_x}(g(z))}_{:= \nabla_{g(z)}^2 \log p(x; g(z))} J_g(z)$$

(exact for relu, leaky-relu in decoder)

Regularization: local geometry

We can further reduce the approximation in terms of the Jacobian of the decoder.

$$\underbrace{H_{f_x}(z)}_{:= \nabla_z^2 \log p(x|z)} \approx J_g(z)^\top \underbrace{H_{p_x}(g(z))}_{:= \nabla_{g(z)}^2 \log p(x; g(z))} J_g(z)$$

(exact for relu, leaky-relu in decoder)

$$\log p_\theta(x|h_\phi(x)) + \frac{1}{2} tr(\boxed{H_{f_x}(h_\phi(x))}\Sigma_{z|x}) - \beta\, KL(q_\phi(z|x) \| p(z))$$

$$\Downarrow$$

$$\log p(x|h(x)) + \frac{1}{2} tr(\boxed{J_g(h(x))^\top H_{p_x}(g(h(x))) J_g(h(x))}\Sigma_{z|x}) - \beta\, KL(q_\phi(z|x) \| p(z))$$

Regularization: local geometry

$$\log p(x|h(x)) + \frac{1}{2}tr(J_g(h(x))^\top \underbrace{H_{p_x}(g(h(x)))}_{:= \nabla^2_{g(z)} \log p(x; g(z))} J_g(h(x))\Sigma_{z|x}) - \beta\, KL(q_\phi(z|x) \| p(z))$$

- $H_{p_x}(g(h(x)))$: Diagonal for pixel-wise independent models
- Minimizes $\left\| [-H_{p_x}(g(h(x)))]^{1/2} J_g(h(x)) \Sigma_{z|x}^{-1/2} \right\|_F^2$

Gaussian $p(z)$ and $q(z|x)$

$$\log p(x|h(x)) + \frac{1}{2} tr(J_g(h(x))^\top \underbrace{H_{p_x}(g(h(x)))}_{:= \nabla^2_{g(z)} \log p(x; g(z))} J_g(h(x)) \Sigma_{z|x}) - \beta \, KL(q_\phi(z|x) \| p(z))$$

For this special case, optimal variational posterior covariance is given by

$$\Sigma_{z|x} = \left(I - \frac{1}{\beta} J_g(h(x))^\top H_{p_x}(g(h(x))) J_g(h(x)) \right)^{-1}$$

Gaussian $p(z)$ and $q(z|x)$

$$\log p(x|h(x)) + \frac{1}{2} tr(J_g(h(x))^\top \underbrace{H_{p_x}(g(h(x)))}_{:= \nabla^2_{g(z)} \log p(x; g(z))} J_g(h(x)) \Sigma_{z|x}) - \beta \, KL(q_\phi(z|x) \| p(z))$$

For this special case, optimal variational posterior covariance is given by

$$\Sigma_{z|x} = \left(I - \frac{1}{\beta} J_g(h(x))^\top \underbrace{H_{p_x}(g(h(x)))} J_g(h(x)) \right)^{-1}$$

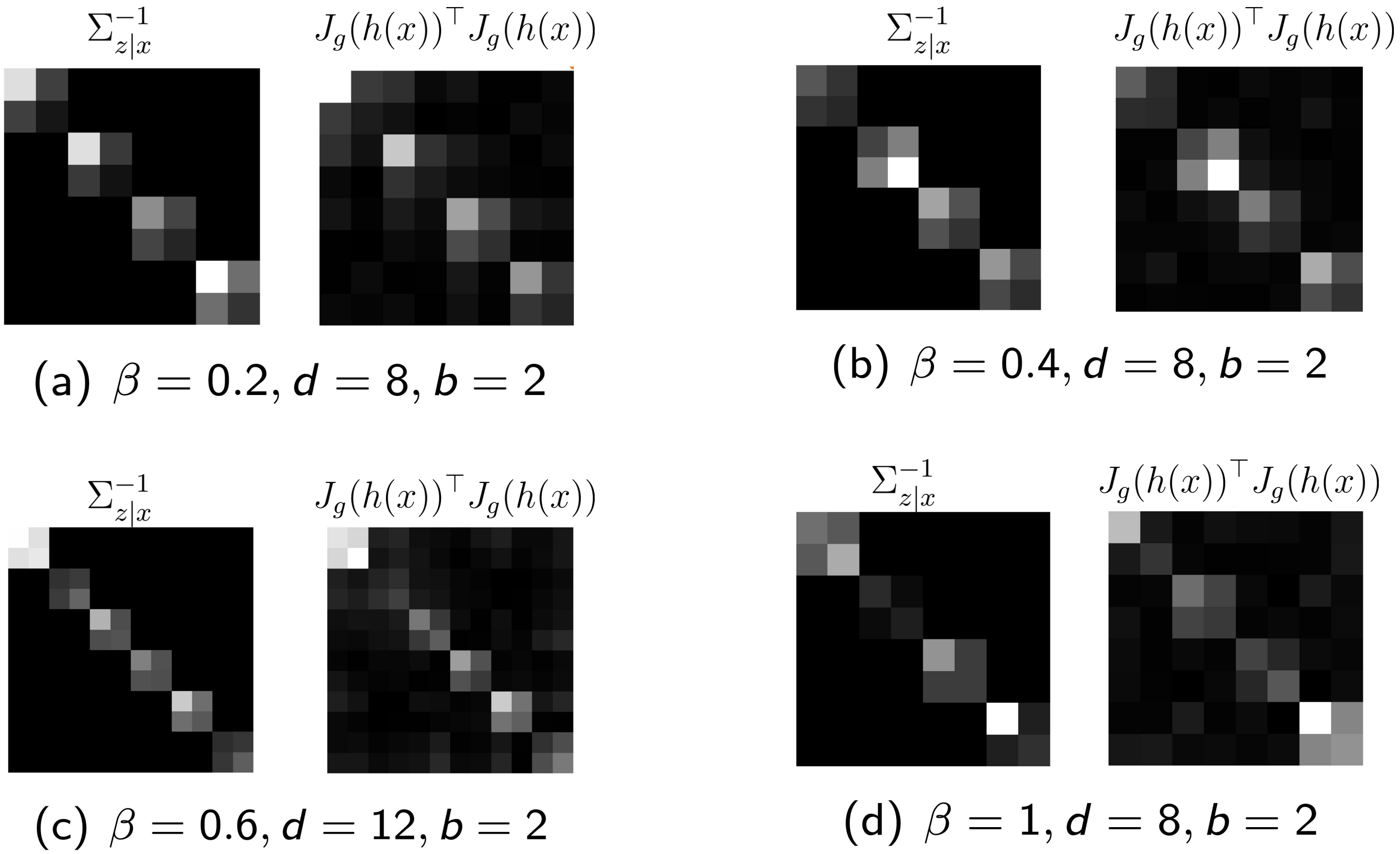
$$= -I \quad \text{for} \quad p(x|z) = N(g(z), I)$$

$$\approx -I \quad \text{for} \quad p(x|z) = \text{Bern}(g(z))$$

A structure on $\Sigma_{z|x}$ influences the Jacobian of the decoder.

More details in the paper about its influence on metric properties of the learned manifold.

MNIST


$\Sigma_{z|x}^{-1}$
$J_g(h(x))^\top J_g(h(x))$
(a) $\beta = 0.2, d = 8, b = 2$
$\Sigma_{z|x}^{-1}$
$J_g(h(x))^\top J_g(h(x))$
(b) $\beta = 0.4, d = 8, b = 2$
$\Sigma_{z|x}^{-1}$
$J_g(h(x))^\top J_g(h(x))$
(c) $\beta = 0.6, d = 12, b = 2$
$\Sigma_{z|x}^{-1}$
$J_g(h(x))^\top J_g(h(x))$
(d) $\beta = 1, d = 8, b = 2$

Gaussian $p(z)$, $q(z|x)$, $p(x|z)$

The β-VAE objective approximates to

GRAE:

$$\min_{g,h} \underbrace{\frac{1}{2}\|x - g(h(x))\|^2}_{\text{Reconstruction error}} + \underbrace{\frac{\beta}{2}\|h(x)\|^2}_{\text{Encoding norm}} + \underbrace{\frac{\beta}{2}\log\left|I + \frac{1}{\beta}J_g(h(x))^\top J_g(h(x))\right|}_{\text{Regularizer on the decoder Jacobian}}$$

Gaussian $p(z)$, $q(z|x)$, $p(x|z)$

The β-VAE objective approximates to

GRAE:

$$\min_{g,h} \underbrace{\frac{1}{2}\|x - g(h(x))\|^2}_{\text{Reconstruction error}} + \underbrace{\frac{\beta}{2}\|h(x)\|^2}_{\text{Encoding norm}} + \underbrace{\frac{\beta}{2}\log\left|I + \frac{1}{\beta}J_g(h(x))^\top J_g(h(x))\right|}_{\text{Regularizer on the decoder Jacobian}}$$

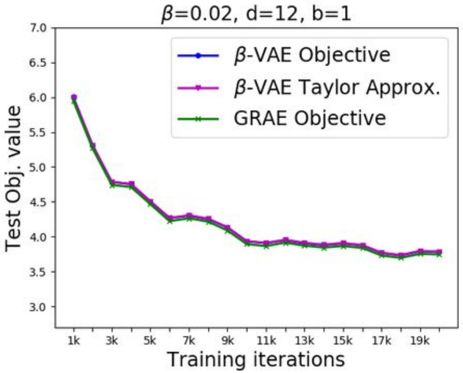
We upper bound the regularizer to make it more tractable:

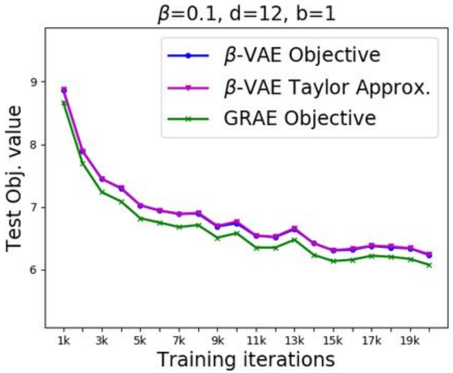
GRAE≈:

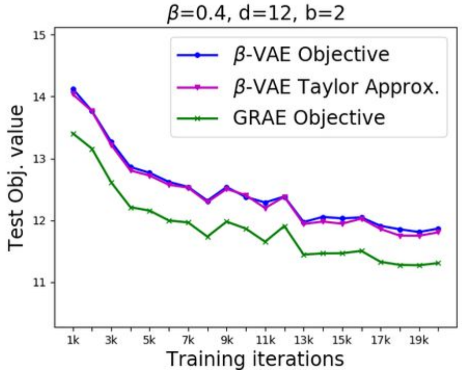
$$\log\left|I + \frac{1}{\beta}J_g(h(x))^\top J_g(h(x))\right| \leq \sum_i \log\left(1 + \frac{1}{\beta}\|[J_g(h(x))]_{:i}\|_2^2\right)$$

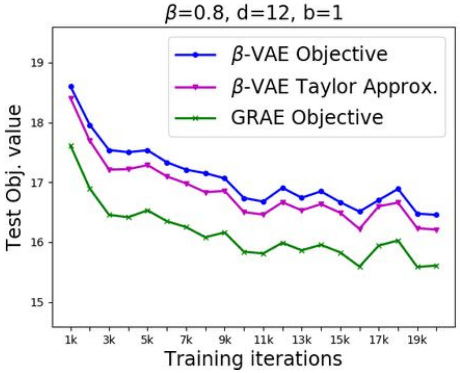
We minimize a stochastic approximation of this upper bound (sampling one column of Jacobian per iteration).

Comparison of objectives

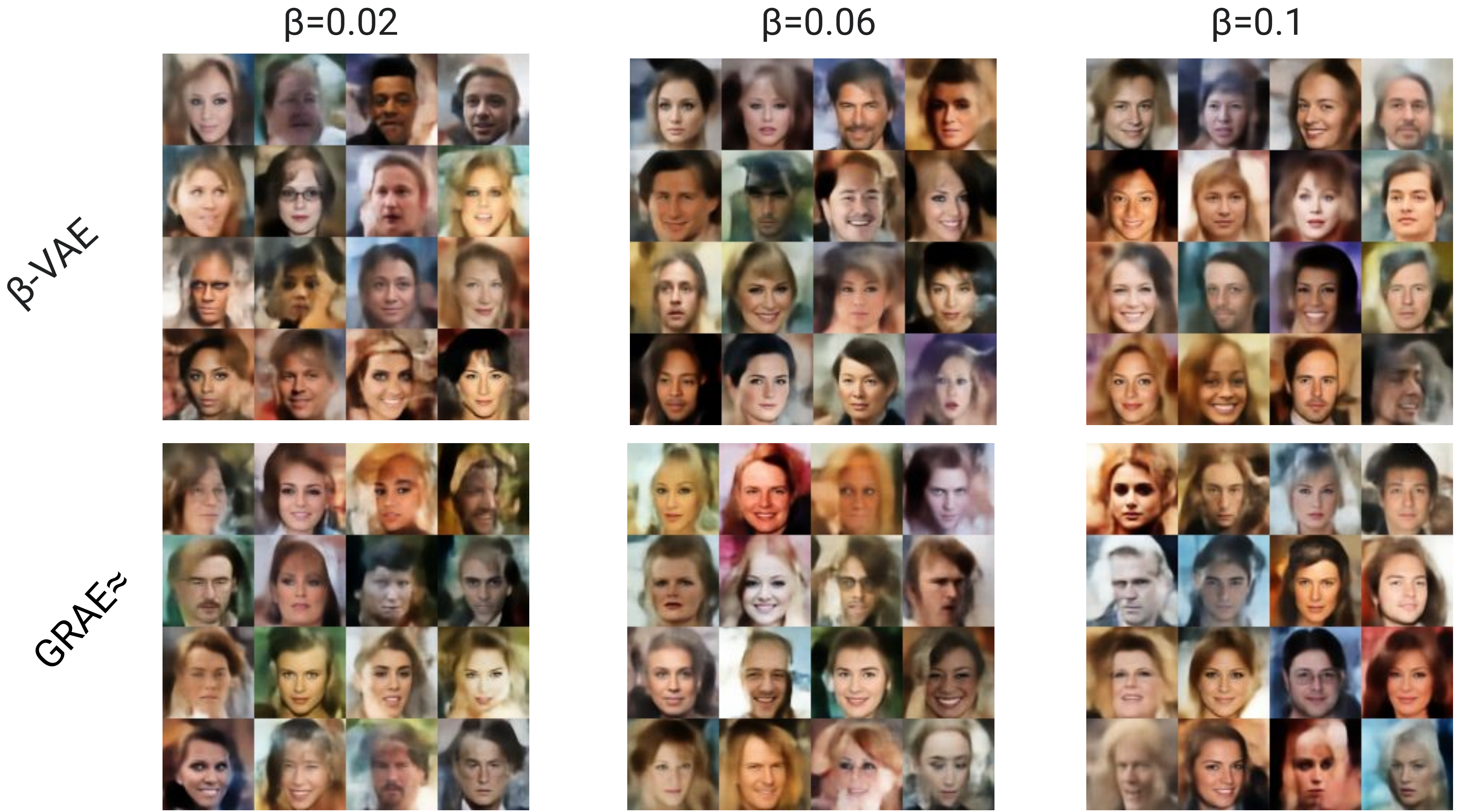








Samples



Samples

β=0.4 β=0.6 β=0.8

β-VAE

GRAE≈

Thanks

For more details: https://arxiv.org/abs/2002.00041