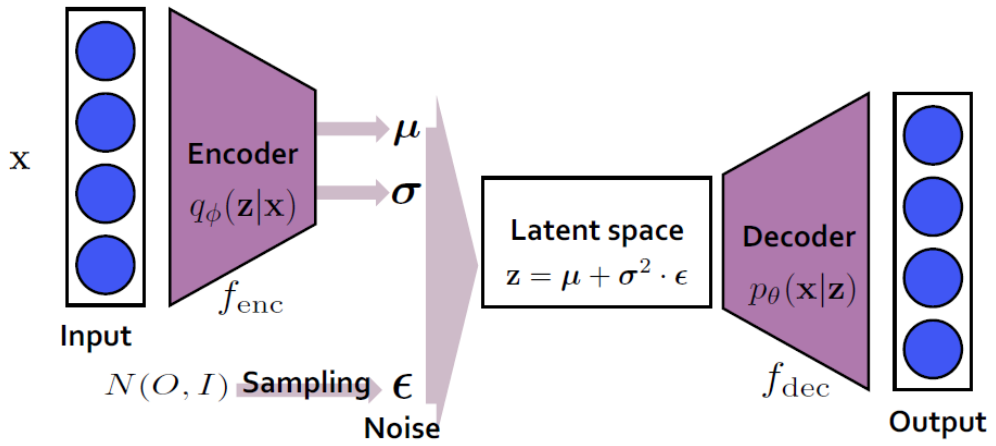
Decoupling β from σ_x in the Gaussian VAE

Presenter: Kim Seung Hwan (overnap@khu.ac.kr)

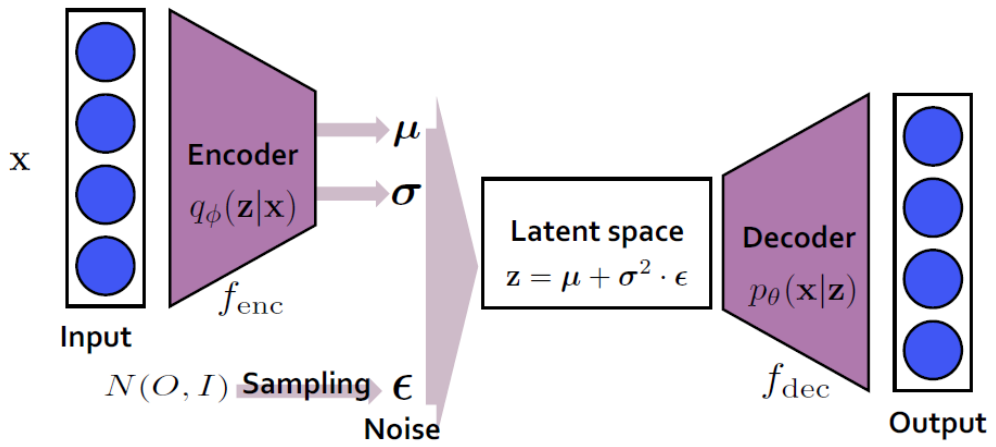
(Gaussian) Variational Autoencoder

- Deep Maximum Likelihood Model
- Decoder $p_\theta(x|z)$ and a $q_\phi(z|x)$ for a Data x and its latent representation z
- $p_\theta(x|z) \sim N(\mu_x(z), \sigma_x^2(z))$, $q_\phi(z|x) \sim N(\mu_z(x), \sigma_z^2(x))$



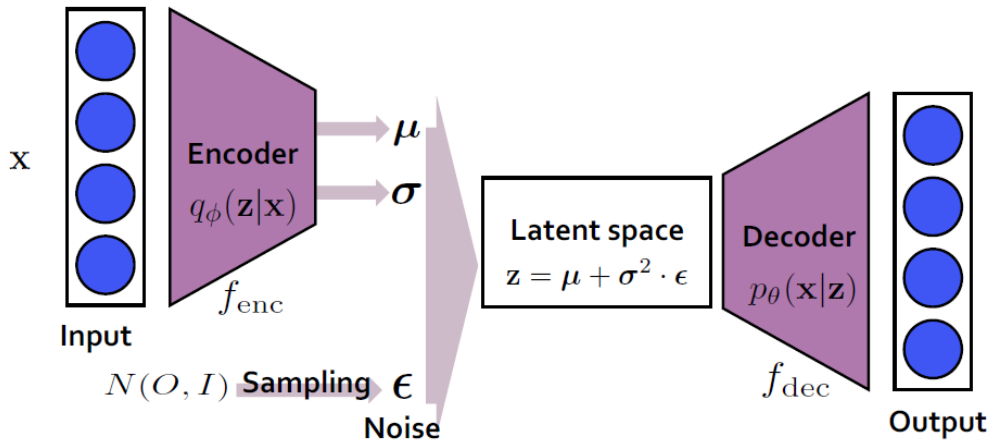
Implementation of Gaussian

- $p_\theta(x|z) \sim N(\mu_x(z), \sigma_x^2(z))$, $q_\phi(z|x) \sim N(\mu_z(x), \sigma_z^2(x))$
- This means that they follow the distributions
- No need to be explicitly sampled!



Implementation of Gaussian (contd.)

- Typical VAEs are implemented with the explicit encoder and implicit decoder
- Specifically, $p_\theta(x|z) \sim N(\mu_x(z), \sigma_x^2(z)I)$ i.e. same variance on all channels
- And $q_\phi(z|x) \sim N(\mu_z(x), diag(\sigma_z^2(x)))$ i.e. diagonal covariance



Evidential Lower Bound

- Optimization target: ELBO (Evidence Lower Bound)

$$-\log p_\theta(\mathbf{x}) \leq E_{\mathbf{z}\sim q_\phi(\mathbf{z}|\mathbf{x})}[-\log p_\theta(\mathbf{x}|\mathbf{z})] + D_{KL}(q_\phi(\mathbf{z}|\mathbf{x})||p_\theta(\mathbf{z}))$$

- If the encoder and decoder are Gaussian,
- $E_{z\sim q_\phi(z|x)}[-\log p_\theta(x|z)] = E_{z\sim q_\phi(z|x^i)}[\frac{(x-\mu_x(z))^2}{2\sigma_x^2(z)} + \frac{1}{2}\log 2\pi\sigma_x^2(z)]$
- $D_{KL}(q_\phi(z|x)||p_\theta(z)) = \frac{1}{2}[|\sigma_z^2(x)| + |\mu_z^2(x)| - |\log \sigma_z^2(x) + 1|]$

ELBO with β

- The β-VAE loss:
- $E_{z\sim q_\phi(z|x)}[-\log p_\theta(x|z)] + \beta D_{KL}(q_\phi(z|x)||p_\theta(z))$

  Reconstruction loss = Distortion Regularization loss = Rate
- This balances the reconstruction and regularization loss
- Each can be interpreted as Rate and Distortion

- This is representative of many improvements to VAE

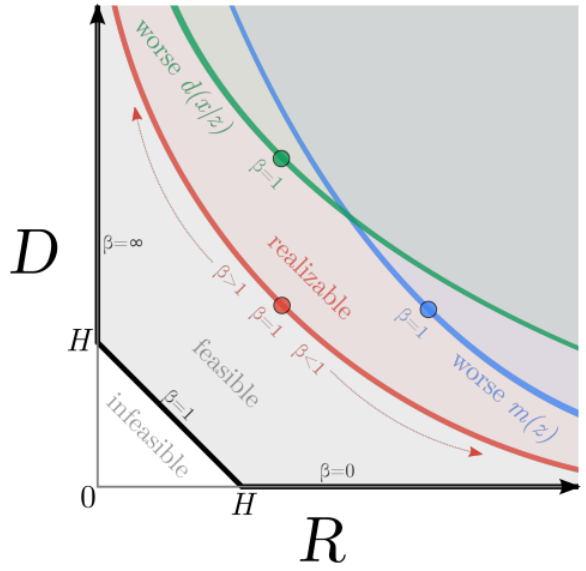


Figure 1. Schematic representation of the phase diagram in the RD-plane. The *distortion* (D) axis measures the reconstruction error of the samples in the training set. The *rate* (R) axis measures the relative KL divergence between the encoder and our own marginal approximation. The thick black lines denote the feasible boundary in the infinite model capacity limit.

The Pitfall When Implementing ELBO

- Practitioner often implement Reconstruction loss as MSE
- And tune a good β value
- $\underbrace{E_{z\sim q_\phi(z|x)}[-\log p_\theta(x|z)]}_{\text{Reconstruction loss = Distortion}} + \beta \underbrace{D_{KL}(q_\phi(z|x)||p_\theta(z))}_{\text{Regularization loss = Rate}}$
- For a Gaussian decoder, the β equals the $2\sigma_x^2(z)$ as a constant
- $E_{z\sim q_\phi(z|x)}[-\log p_\theta(x|z)] = E_{z\sim q_\phi(z|x^i)}[\frac{(x-\mu_x(z))^2}{2\sigma_x^2(z)} + \frac{1}{2}\log 2\pi\sigma_x^2(z)]$
- This seems like a good alternative interpretation,
- There are 2 problems though

Problem 1. *Learnable σ_x is important*

- The first problem is that the constant σ_x itself is pathological

- Dai, Bin, Li Wenliang, and David Wipf. "On the value of infinite gradients in variational autoencoder models." Advances in Neural Information Processing Systems 34 (2021): 7180-7192.
- Let us look at the argument of the above paper!

Optimal Sparse Representation

- The minimum information that can represent the data manifold perfectly
- In the words of the authors, “the most parsimonious latent representation”

Definition 1 *An autoencoder-based architecture (VAE or otherwise) with decoder* $\boldsymbol{\mu}_x(\cdot;\theta)$*, constraint* $\theta \in \Theta$*, and arbitrary encoder* $\boldsymbol{\mu}_z$ *component*[1] *produces an* ***optimal sparse representation*** *of a training set* $\boldsymbol{X}$ *w.r.t.* Θ *if the following two conditions simultaneously hold:*

(i) The reconstruction error is zero, meaning

$$\frac{1}{n}\sum_{i=1}^{n}\left\|\boldsymbol{x}^{(i)} - \boldsymbol{\mu}_x\left[\boldsymbol{\mu}_z\left(\boldsymbol{x}^{(i)};\phi\right);\theta\right]\right\|_2^2 = 0. \tag{4}$$

(ii) Conditioned on achieving perfect reconstructions per criteria (i) above, the number of latent dimensions such that $\mu_z\left(\boldsymbol{x}^{(i)};\phi\right)_j = 0$ *for all* i *is maximal across any* $\theta \in \Theta$ *and any encoder function* $\boldsymbol{\mu}_z$*. A* j*-th latent dimension so-defined provides no benefit in reducing the reconstruction error and could in principle be removed from the model.*

Why Mean?

- Is it okay to consider only the mean (of the encoder and decoder?)
- Dai, Bin, and David Wipf. "Diagnosing and Enhancing VAE Models." International Conference on Learning Representations. 2018.

$$2\mathbb{KL}\left[q_\phi(\boldsymbol{z}|\boldsymbol{x})||p(\boldsymbol{z})\right] \equiv \text{trace}\left[\boldsymbol{\Sigma}_z\right] + \|\boldsymbol{\mu}_z\|_2^2 - \log|\boldsymbol{\Sigma}_z| \approx -\hat{r}\log\gamma + O(1).$$

$\gamma = \sigma_x^2(z)$

Estimated low-noise latent dimensions ≈ Informative dimension

Power

↓

Result

Therefore, *in the neighborhood of optimal solutions the VAE will naturally seek to produce perfect reconstructions using the fewest number of clean, low-noise latent dimensions*,

to Definition 1. More concretely, for unneeded latent dimensions the posterior is pushed to the prior to optimize the KL regularizer, i.e., $q_\phi\left(z_j|\boldsymbol{x}^{(i)}\right) = \mathcal{N}(0,1)$ for all i, which amounts to uninformative noise that will be filtered by the decoder so as not to impact reconstructions. In contrast, for informative dimensions the posterior variance satisfies $\sigma_z\left(\boldsymbol{x}^{(i)};\phi\right)_j \to 0$ for all i. Collectively, this

Why Sparse?

- Why is sparsity necessary for downstream task e.g. generation?
- The authors explain with an example of inlier-outlier
- I am still confused; would not it be better to use all channels for generation?

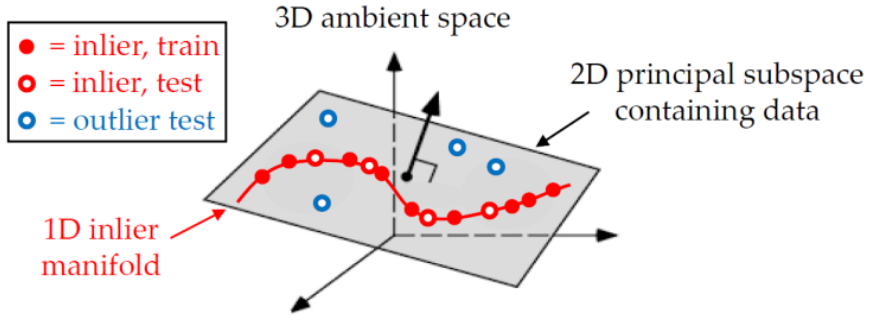


Figure 1: The importance of optimal sparse representations in screening outliers. In this example, the simple 2D principal subspace obtainable by PCA can perfectly reconstruct the inlier manifold shown in red. But this requires using two separate informative dimensions, allowing both inliers *and* outliers to be reconstructed with zero error within this subspace. In contrast, it is only by recovering the curved 1D inlier manifold, which relies on a single informative dimension, that inliers and outliers can be differentiated. Please see supplementary for practical example using real data.

Infinite Gradient Is Integral

- For generalized loss of an autoencoder-like model, (i.e. MSE + Regularizing-Z)

$$\mathcal{L}_{g,h}(\theta,\phi) \triangleq g\left(\frac{1}{dn}\sum_{i=1}^{n}\left\|\boldsymbol{x}^{(i)}-\boldsymbol{\mu}_x\left(\boldsymbol{z}^{(i)};\theta\right)\right\|_2^2\right)+\frac{1}{d}\sum_{k=1}^{\kappa}h\left(\frac{1}{n}\|\boldsymbol{z}_k\|_2^2\right),$$
$$\text{s.t. } \boldsymbol{z}^{(i)}=\boldsymbol{\mu}_z\left(\boldsymbol{x}^{(i)};\phi\right)\ \forall i,\ \theta\in\Theta,$$

- Minimizing the loss with bounded g, h (which means a finite gradient)
- *Cannot* produce perfect reconstruction or optimal sparsity

Theorem 4 *For any functions $g:\mathbb{R}^+\to\mathbb{R}$ and $h:\mathbb{R}^+\to\mathbb{R}$ with bounded gradients, and any dimension set $\{d,\kappa,r\}$ that order as $d\geq\kappa>r>0$, there exists data $\boldsymbol{X}=\{\boldsymbol{x}^{(i)}\}_{i=1}^n\in\mathbb{R}^{d\times n}$ and decoder $\{\boldsymbol{\mu}_x(\boldsymbol{z};\theta),\theta\in\Theta\}$ (with the capacity to reconstruct $\boldsymbol{x}$ lying within some parameterized family of κ-dimensional manifolds) which satisfy the following:*

(a) $\frac{1}{n}\sum_{i=1}^n\left\|\boldsymbol{x}^{(i)}-\boldsymbol{\mu}_x\left[\boldsymbol{z}^{(i)};\theta\right]\right\|_2^2=0$ *for some $\theta\in\Theta$ and $\boldsymbol{Z}\in\mathbb{R}^{\kappa\times n}$ with $\|\boldsymbol{z}_k\|_2>0$ for r rows and zero elsewhere.*

(b) Minimizing $\mathcal{L}_{g,h}(\theta,\phi)$ over θ and any possible encoder produces either a solution with $\frac{1}{n}\sum_{i=1}^n\left\|\boldsymbol{x}^{(i)}-\boldsymbol{\mu}_x\left[\boldsymbol{z}^{(i)};\theta\right]\right\|_2^2>0$ (i.e., imperfect reconstruction), or one where $\|\boldsymbol{z}_k\|_2>0$ for strictly more than r rows of $\boldsymbol{Z}$ (i.e., not maximally sparse).

Infinite Gradient Is Integral (contd.)

- An infinite gradient is thus a necessary condition (not a sufficient condition)
- Even with other AEs, infinite gradient are essential e.g. AAE, WAE, DAE
- The often-ignored decoder variance $\sigma_x^2(z)$ of VAEs is also essential!
- Now let us look at it experimentally

Greens can be infinitely large or small

Reconstruction loss (Distortion)

$$\mathcal{L}(\theta,\phi) \equiv \frac{1}{n}\sum_{i=1}^{n}\left\{\mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x}^{(i)})}\left[\frac{1}{\gamma}\|\boldsymbol{x}^{(i)}-\boldsymbol{\mu}_x(\boldsymbol{z};\theta)\|_2^2\right]+d\log\gamma\right. \tag{3}$$

$$\left.+\left\|\boldsymbol{\sigma}_z\left(\boldsymbol{x}^{(i)};\phi\right)\right\|_2^2-\log\left|\operatorname{diag}\left[\boldsymbol{\sigma}_z\left(\boldsymbol{x}^{(i)};\phi\right)\right]^2\right|+\left\|\boldsymbol{\mu}_z\left(\boldsymbol{x}^{(i)};\phi\right)\right\|_2^2\right\}.$$

Regularization loss (Rate)

Visualization

- A large σ_x^2 smooths the loss, and a small σ_x^2 reveals the exact optimal
- So if one knows the optimal σ_x^2 (≈ 0) and fixes the σ_x^2 from the start,
- One might get the bad result (note the constant σ_x^2 model $\approx \beta$-VAE)

	CelebA	
	Rec. Err.	MMD
Learnable γ	352.8	93.3
Fix $\gamma = \gamma^*$	349.9	291.8

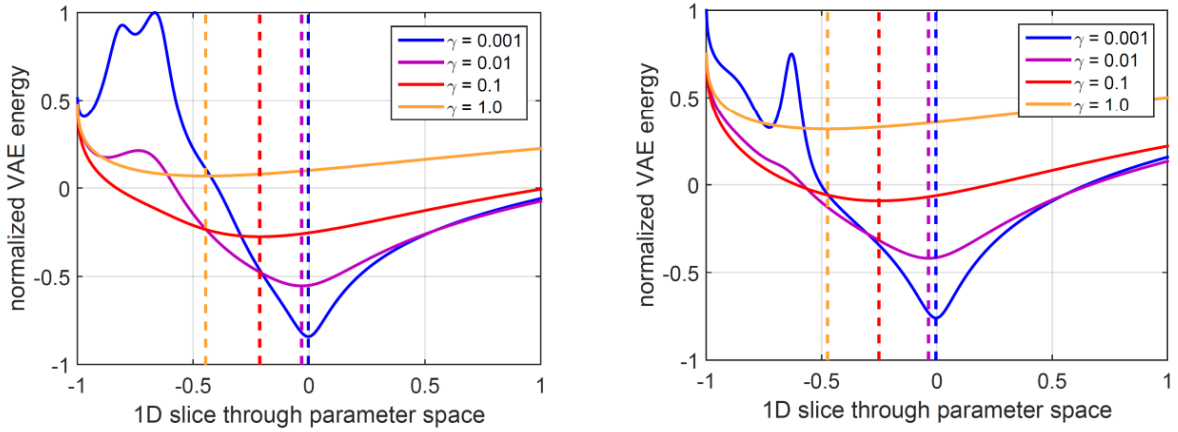


Figure 3: Plots (a) and (b) show two sets of representative 1D slices through the VAE objective function (3) as the value of γ is varied. Dashed vertical lines indicate the x-axis location of the minimal value of each respective slice and γ setting. And for both plots (a) and (b) the 1D slices are set such that an optimal sparse representation would occur at zero on the x-axis when $\gamma \to 0$. It can be observed that disconnected local minima only occur when γ is small.

Calibrated Decoder

- But the instability of the infinite gradient needs to be addressed!
- Rybkin, Oleh, Kostas Daniilidis, and Sergey Levine. "Simple and effective VAE training with calibrated decoders." International Conference on Machine Learning. PMLR, 2021.

	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	$<$ **6359**	71.5	$<$ 9179	46.13	$<$ **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	$<$ **9085**	43.19	$<$ 8443	143.1	$<$ **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	$<$ **-3659**	70.4	$<$ **-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	$<$ **-8446**	**60.3**	<-3333	27.25	<-5677	**101.4**	$<$ **-14173**	34.13
Opt. per-image σ-VAE		66.01		26.28		**104.0**		**33.21**

Problem 2. *β was designed independently*

- The second problem is that the β was designed independently
- Of the σ_x and more importantly, of the Gaussian VAE!

- Higgins, Irina, et al. "beta-vae: Learning basic visual concepts with a constrained variational framework." International conference on learning representations. 2016.

$$\mathcal{F}(\theta, \phi, \beta; \mathbf{x}, \mathbf{z}) \geq \mathcal{L}(\theta, \phi; \mathbf{x}, \mathbf{z}, \beta) = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})] - \beta \, D_{KL}(q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z}))$$

- It is not made to target Gaussian VAEs

β is completely new term!

Alemi, Alexander, et al. "Fixing a broken ELBO." International conference on machine learning. PMLR, 2018.

$$\min_{e(z|x),m(z),d(x|z)} \int dx\, p^*(x) \int dz\, e(z|x) \left[-\log d(x|z) + \beta \log \frac{e(z|x)}{m(z)} \right].$$

(stipulate β as a new term independent of σ_x)

Lucas, James, et al. "Don't blame the elbo! a linear vae perspective on posterior collapse." Advances in Neural Information Processing Systems 32 (2019).

"Importantly, the Gaussian partition function for a Gaussian observation model (the last term on the RHS of Eq. (10)) prevents ELBO from deviating from the β-VAE's objective with a β-weighted KL term while maintaining the benefits to representation learning when σ2 is small.

Burgess, Christopher P., et al. "Understanding disentangling in $\beta $-VAE." arXiv preprint arXiv:1804.03599 (2018).

$$\mathcal{L}(\theta, \phi; \mathbf{x}, \mathbf{z}, \beta) = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})] - \beta\, D_{KL}\big(q_\phi(\mathbf{z}|\mathbf{x}) \parallel p(\mathbf{z})\big)$$

(same as the first above)

β can be viewed as the constant σ_x!

Dai, Bin, Li Wenliang, and David Wipf. "On the value of infinite gradients in variational autoencoder models." Advances in Neural Information Processing Systems 34 (2021): 7180-7192.

"For Gaussian VAE models (which is our focus), this scale factor effectively makes no difference if a fixed decoder variance is adopted. In this situation, β can just be directly absorbed into γ, and the d log γ normalization factor from (3) can be viewed as an irrelevant constant."

Rybkin, Oleh, Kostas Daniilidis, and Sergey Levine. "Simple and effective VAE training with calibrated decoders." International Conference on Machine Learning. PMLR, 2021.

"The β-VAE objective is then equivalent to a σ-VAE with a constant variance σ^2=β/2 (for a particular learning rate setting.)"

Conclusion

- In conclusion, they are equal as formulas (from the perspective of optimization)
- However, assuming a constant-variance decoder is problematic (for performance)
- And many of the β-VAE analyzes interpreted β
- As a separated parameter from σ_x (from the perspective of information theory)

- Therefore, It is better to research and/or implement them in isolation!

Method: *Decoupling σ_x from β*

- Adopt learnable σ_x or analytically optimal $\sigma_x^2 = MSE$
- Perhaps this would be more like the best of the existing β-VAEs
- Rather than a great performance boost (exhibited in the calibrated decoder paper)

- Since I consider the β to be completely different from the σ_x^2,
- Employ β again to this log-σ-VAE
- And observe the RD-curve or disentanglement,
- Which is a representative effect of β

Method: *Decoupling σ_x from β* (contd.)

- Apply this decoupling to the previous β-input VAE
- I hope this will restore a more proper RD-curve
- 'Proper' means corresponding to studies like "Fixing a broken ELBO" paper etc.

- This solves the gap between the theory and practice,
- And allows σ_x^2 and β to be used for their designed purpose

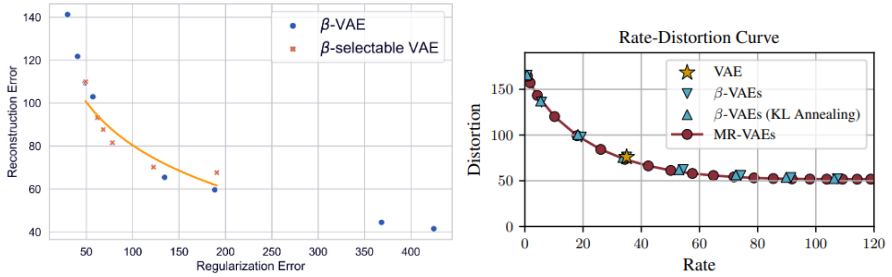


Figure 3: Reconstruction and regularization errors of ELBO for various βs. For ELBO, i.e. the learning objective, a single β-Selectable VAE is capable of approximate many β-VAEs. This graph can be interpreted as a rate-distortion curve.

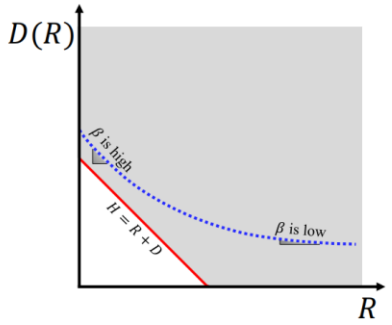


Figure 2: Schematic view of distortion-rate function. A red line corresponds to the theoretical lower bound of the rate and distortion. By varying β of β-VAE, we could achieve the points on a blue dashed curve, the sub-optimal distortion-rate function, which is best achievable with VAEs.

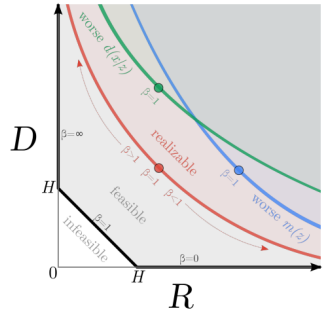


Figure 1. Schematic representation of the phase diagram in the RD-plane. The *distortion* (D) axis measures the reconstruction error of the samples in the training set. The *rate* (R) axis measures the relative KL divergence between the encoder and our own marginal approximation. The thick black lines denote the feasible boundary in the infinite model capacity limit.

Additional Methods

1. Decompose $D_{KL}(q_\phi(z|x)||p(z))$ *(from second direction)*
   - When strengthening the regularization term,
   - We do not want to suffer the performance penalty while taking advantage of good sampling
2. Make sampler *(from third direction)*
   - The posterior sampler of a VAE is often easily implemented as a *second VAE*
   - Can one β-input VAE perform complete sampling of other β-input VAE?

Aggregated Posterior

- If this is theoretically true, RD-Curve will not be able to express all of VAEs
- A new axis for some "sampling ease" should be added
- The candidate emerges from the decomposition mentioned earlier
- This is the ***second research direction***

Table 1: FID scores of VAEs for both

	β (value or range)	CelebA
β-VAE	1.0	65.33
β-VAE	0.1	57.19
β-VAE	0.01	63.42
Ours	1.0	58.41
Ours	0.1	55.43
Ours	0.01	55.28

$$\mathcal{L}_\beta(\phi,\theta) = \underbrace{\mathbb{E}_{p_d(\mathbf{x})}[\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[-\log p_\theta(\mathbf{x}|\mathbf{z})]}_{\text{Distortion}} + \beta \underbrace{\mathbb{E}_{p_d(\mathbf{x})}[D_{\mathrm{KL}}(q_\phi(\mathbf{z}|\mathbf{x}), p(\mathbf{z}))]}_{\text{Rate}},$$

$$\mathcal{L}(\theta,\phi) = \mathbb{E}_{q_\phi(z,x)}\left[\log \frac{p_\theta(x,z)}{p_\theta(x)p(z)} + \log \frac{q_\phi(z)q(x)}{q_\phi(z,x)} + \log \frac{p_\theta(x)}{q(x)} + \log \frac{p(z)}{q_\phi(z)}\right],$$

$$= \mathbb{E}_{q_\phi(z,x)}\left[\underbrace{\log \frac{p_\theta(x\mid z)}{p_\theta(x)}}_{①} - \underbrace{\log \frac{q_\phi(z\mid x)}{q_\phi(z)}}_{②}\right] - \underbrace{\mathrm{KL}(q(x)\,\|\,p_\theta(x))}_{③} - \underbrace{\mathrm{KL}(q_\phi(z)\,\|\,p(z))}_{④}.$$

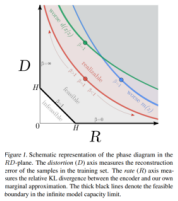
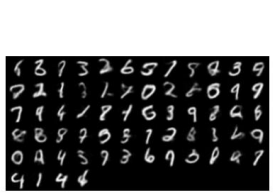


Figure 1. Schematic representation of the phase diagram in the RD-plane. The *distortion* (D) axis measures the reconstruction error of the samples in the training set. The *rate* (R) axis measures the relative KL divergence between the encoder and our own marginal approximation. The thick black lines denote the feasible boundary in the infinite model capacity limit.

Alemi, Alexander, et al. "Fixing a broken ELBO." International conference on machine learning. PMLR, 2018.
Burgess, Christopher P., et al. "Understanding disentangling in β-VAE." arXiv preprint arXiv:1804.03599 (2018).

Aggregated Posterior

- I might find a practical use for the observation: the posteriors are aligned
- The recursive heuristic I tried would be one of them (albeit with bad results)
- It is an interesting property, but I do not know how to utilize it
- This may be the third research direction



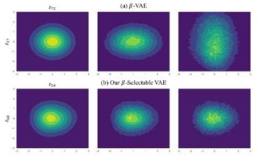


Figure 2: Aggregated posteriors for arbitrarily chosen latent channels. Each row shares a model and channels, and distributions are placed in order of decreasing β. The posteriors of our model (b) are *aligned* relative to their counterparts (a).

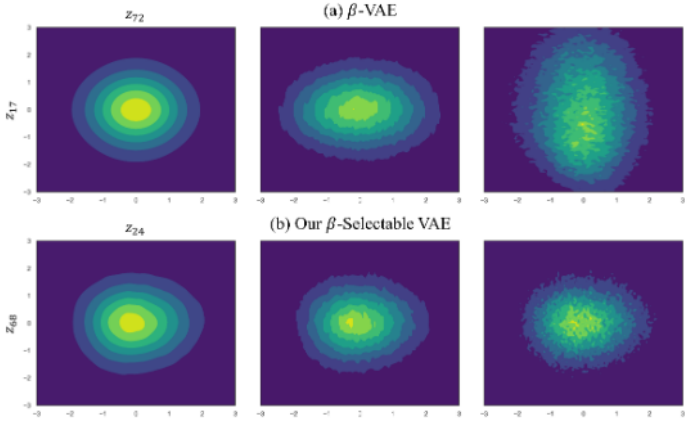
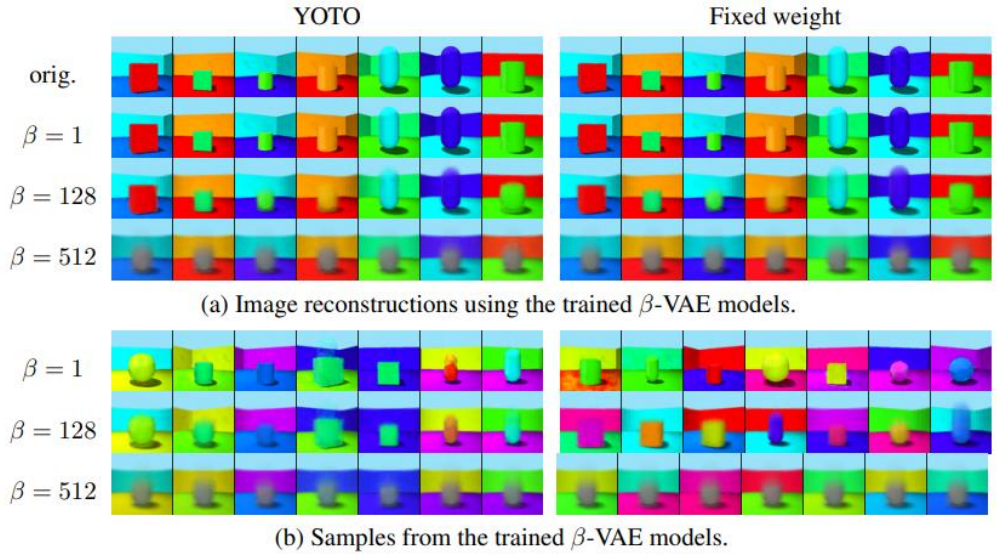


Figure 2: Aggregated posteriors for arbitrarily chosen latent channels. Each row shares a model and channels, and distributions are placed in order of decreasing β. The posteriors of our model (b) are *aligned* relative to their counterparts (a).

Previous Works of β-input strategy

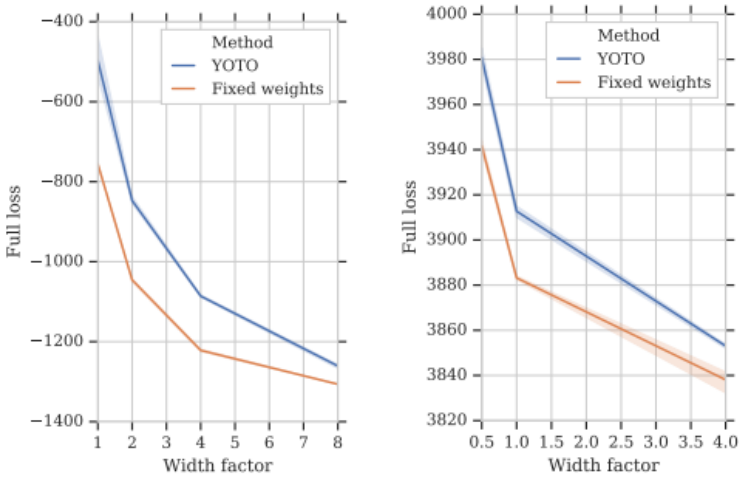
1. YOTO: You Only Train Once

- A naïve approach to approximating linear-weighted loss with one Neural Network
- Possibly the first study to make VAE's β modifiable (input-able)
- Dosovitskiy, Alexey, and Josip Djolonga. "You only train once: Loss-conditional training of deep networks." International conference on learning representations. 2019.



(a) Image reconstructions using the trained β-VAE models.

(b) Samples from the trained β-VAE models.



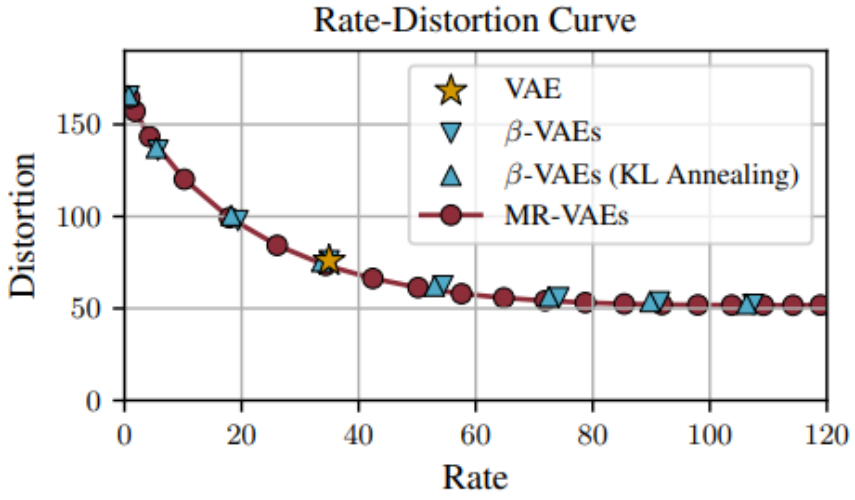
(a) CIFAR-10.

(b) Shapes3D.

Previous Works of β-input strategy (contd.)

2. Multi-Rate VAE

- Approach with hypernetwork methodology
- Analysis mainly with RD-curve
- Bae, Juhan, et al. "Multi-Rate VAE: Train Once, Get the Full Rate-Distortion Curve." The Eleventh International Conference on Learning Representations. 2023.



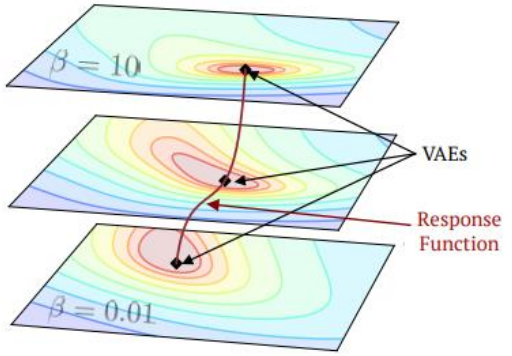


Figure 3: Instead of training several VAEs for each desired KL weight β, MR-VAEs learn the response functions with a hypernetwork in a single training run.

Discussion

- Optimal (minimal) sparse representation is necessary for generation?
- Conversely, is not a maximal dense representation (bounded by latent dimension)
- Required empirically?

- While the authors' argument makes intuitive sense,
- It conflicts with my previous understanding of generation…

Why Sparse?

- Why is sparsity necessary for downstream task e.g. generation?
- The authors explain with an example of inlier-outlier
- I am still confused; would not it be better to use all channels for generation?

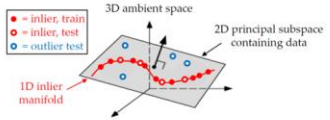


Figure 1: The importance of optimal sparse representations in screening outliers. In this example, the simple 2D principal subspace obtainable by PCA can perfectly reconstruct the inlier manifold shown in red. But this requires using two separate informative dimensions, allowing both inliers *and* outliers to be reconstructed with zero error within this subspace. In contrast, it is only by recovering the curved 1D inlier manifold, which relies on a single informative dimension, that inliers and outliers can be differentiated. Please see supplementary for practical example using real data.

Thank you for listening

Presenter: Kim Seung Hwan (overnap@khu.ac.kr)