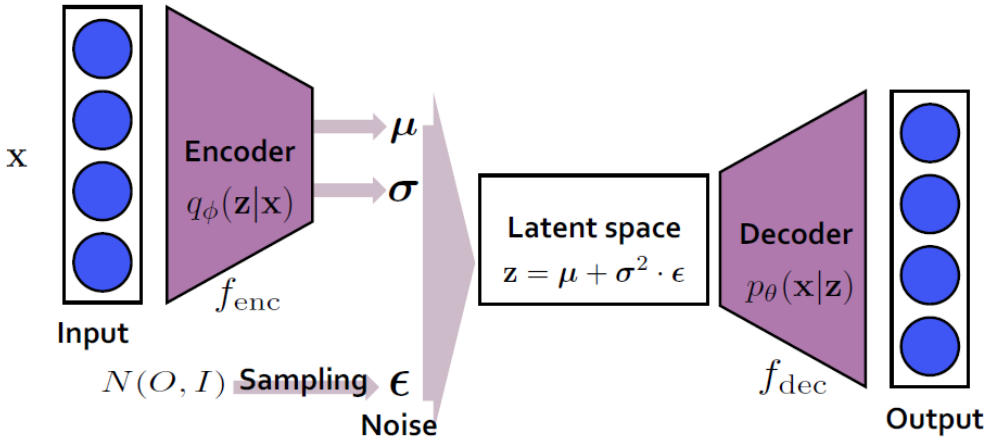
Optimal Sparse Representation & Infinite Gradient

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

Yet Another VAE PT

- 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.



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

Problem & Definition

Suppose we have access to continuous variables $\boldsymbol{x} \in \boldsymbol{\chi}$ that are drawn from ground-truth measure μ_{gt}. This measure assigns probability mass $\mu_{gt}(d\boldsymbol{x})$ to the infinitesimal $d\boldsymbol{x}$ residing within $\boldsymbol{\chi} \subset \mathbb{R}^d$ such that we have $\int_{\boldsymbol{\chi}} \mu_{gt}(d\boldsymbol{x}) = 1$. This formalism allows us to consider data that may lie on or near an r-dimensional manifold embedded in $\mathbb{R}^d$ (implying $r < d$), capturing the notion of low-dimensional structure relative to the high-dimensional ambient space.

Because of the possibility of an unknown latent manifold, it is common to approximate the corresponding ground-truth measure via a density model parameterized as

$$p_\theta(\boldsymbol{x}) = \int p_\theta(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{z})d\boldsymbol{z}. \tag{1}$$

$$q_\phi(\boldsymbol{z}|\boldsymbol{x}) = \mathcal{N}(\boldsymbol{z}|\boldsymbol{\mu}_z, \mathrm{diag}[\boldsymbol{\sigma}_z]^2)$$
$$p_\theta(\boldsymbol{x}|\boldsymbol{z}) = \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_x, \gamma \boldsymbol{I})$$

Reconstruction loss (Distortion)

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

Regularization loss (Rate)

Optimal Sparse Representation

- The minimum information that can represent the data manifold
- 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.

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

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?

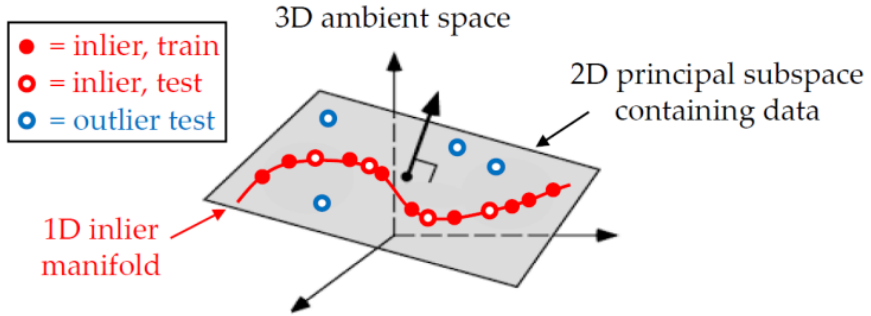


Figure 1: The importance of optimal sparse representations in screening outliers. In this example, the simple 2D principal subspace obtainable by PCA can perfectly reconstruct the inlier manifold shown in red. But this requires using two separate informative dimensions, allowing both inliers *and* outliers 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 + Regularize-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 unbounded 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

- 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 $\gamma = \sigma_x(z)$ 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 γ smooths the loss, and a small γ reveals the exact optimal
- So if one knows the optimal γ (near 0) and fixes the γ from the start,
- One might get the bad result (note the constant γ model $\approx \beta$-VAE)

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

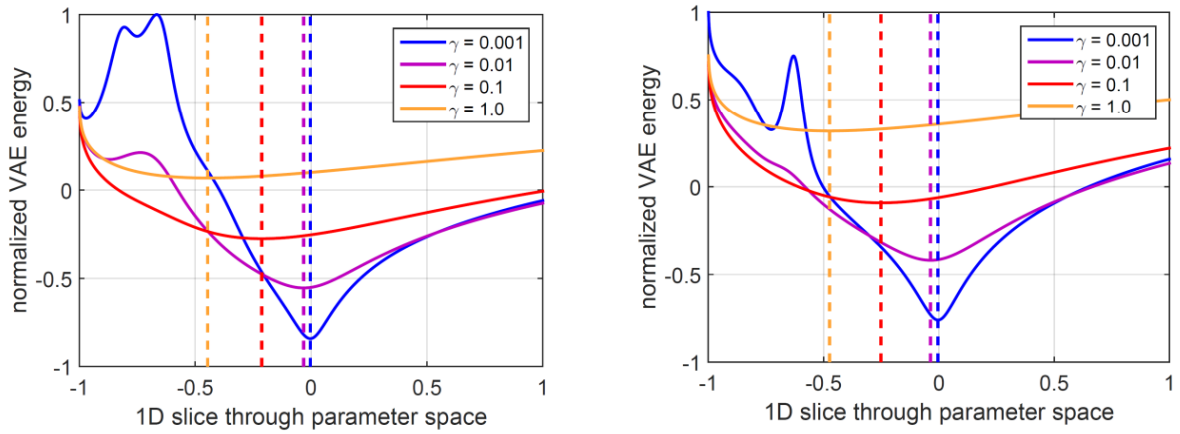


Figure 3: Plots (a) and (b) show two sets of representative 1D slices through the VAE objective function (3) as the value of γ is varied. Dashed vertical lines indicate the x-axis location of the minimal value of each respective slice and γ setting. And for both plots (a) and (b) the 1D slices are set 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**

Decoupling σ_x and β

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.)"

Decoupling σ_x and β

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)

Decoupling σ_x and β

- 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)

- It is therefore better to research and/or implement this in isolation!

Decoupling σ_x and β

- I plan to apply this decoupling to my β-input VAE
- In a rough experiment, the results are interesting…

- 3 contemplated experiments
    1. RD curve of β-VAE and decoupled β-VAE
    2. Check if the decoupled β-input VAE performs better than previous works
    3. Make a β-input sampler (the sampler of a VAE is easily implemented as a second VAE)

- I am sure of the motive, but worried that the novelty will be lacking
- Because the implementation is too bland! (just *log* to *MSE* from previous works)

Discussion

- Optimal (minimal) sparse representation is necessary for generation?
- Conversely, is not a maximal sparse 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?

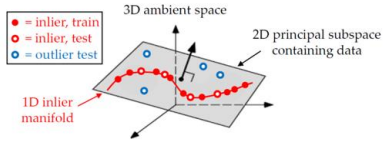


Figure 1: The importance of optimal sparse representations in screening outliers. In this example, the simple 2D principal subspace obtainable by PCA can perfectly reconstruct the inlier manifold shown in red. But this requires using two separate informative dimensions, allowing both inliers *and* outliers 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)