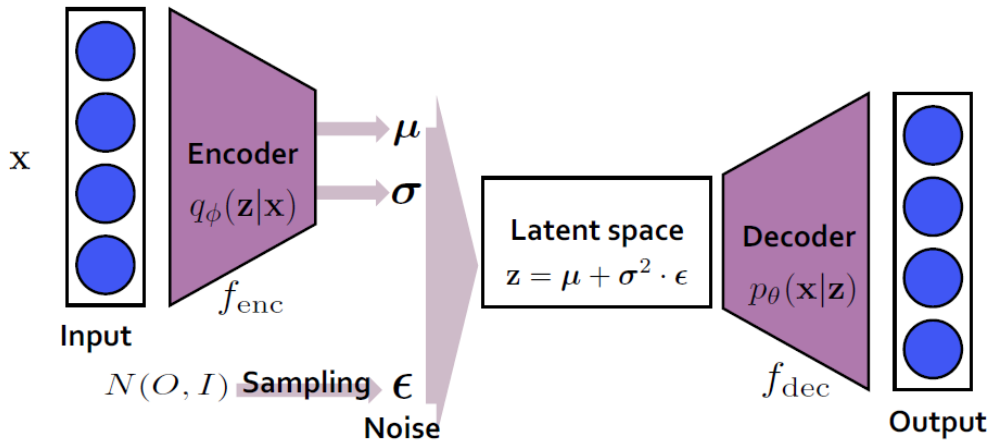
VAE:
Rate-Distortion Curve & Aggregated Posterior

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

What is the Variational Autoencoder?

- Deep Maximum Likelihood Model
- Deep Latent Generative Model
- An autoencoder with gaussian noise regularization (with some fine math support!)
- A Decoder $p_\theta(x|z)$ and a $q_\phi(z|x)$ for a Data x and its latent representation z



Kingma, Diederik P., and Max Welling. "Auto-encoding variational bayes." *arXiv preprint arXiv:1312.6114* (2013).

Evidential Lower Bound?

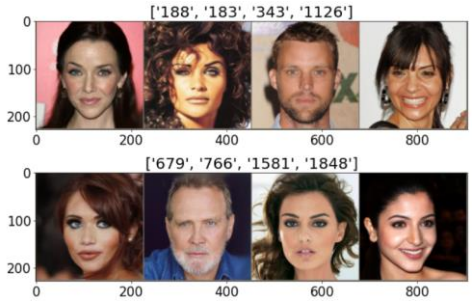
- Optimization target: ELBO (Evidence Lower Bound)
- ELBO consists of the reconstruction term and regularization term

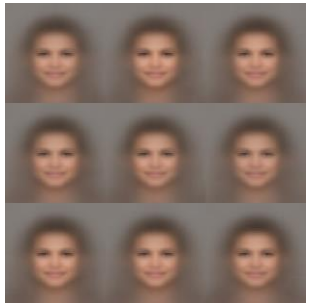
$$-\log p_\theta(\mathbf{x}) \leq E_{\mathbf{z}\sim q_\phi(\mathbf{z}|\mathbf{x})}[-\log p_\theta(\mathbf{x}|\mathbf{z})] + D_{KL}(q_\phi(\mathbf{z}|\mathbf{x})||p_\theta(\mathbf{z}))$$

- Since ELBO is the bound of the negative log-likelihood,
- VAE can be viewed as a deep maximum likelihood model

Pros and Cons of VAE

- Pros
  - Solid mathematical background
  - Lightweight; simple structure and implementation (compared to the Diffusion)
  - No need adversarial strategy (compared to the GAN)
  - Low-dimensional latent variable
- Cons
  - Posterior collapse (autodecoding-like behavior – always outputting the same)
  - Blurry output (bad reconstruction)
  - Poor sampling quality (samples from prior are noticeably worse than reconstruction)





Blame the ELBO

- Many VAE enhancements fix ELBO
- β-VAE

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

- And…

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

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

Paper	Objective
Kingma and Welling [2013], Rezende et al. [2014]	① + ② + ③ + ④
Higgins et al. [2016]	① + ③ + β (② + ④)
Kumar et al. [2017]	① + ② + ③ + λ ④
Zhao et al. [2017]	① + ③ + λ ④
Alemi et al. [2018], Burgess et al. [2018]	① + ③ + γ|(② + ④) − C|
Gao et al. [2018]	① + ② + ③ + ④ − λ ②$^{\mathrm{a}}$
Achille and Soatto [2018]	① + ③ + β ② + γ Ⓐ*
Kim and Mnih [2018], Chen et al. [2018]	① + ② + ③ + Ⓑ + β Ⓐ*
HFVAE (this paper)	① + ③ + ⓘⓘ + α ② + β Ⓐ + γ ①

Table 1: Comparison of objectives in autoencoding deep generative models. The asterisk Ⓐ* indicates that the prior factorizes, i.e. $p(\boldsymbol{z}) = \prod_d p(\boldsymbol{z}_d)$. The notation ②$^{\mathrm{a}}$ refers to restriction of the mutual information ② to a subset of "Anchor" variables $\boldsymbol{z}_a$.

Higgins, Irina, et al. "beta-vae: Learning basic visual concepts with a constrained variational framework." International conference on learning representations. 2016.
Burgess, Christopher P., et al. "Understanding disentangling in $\beta $-VAE." arXiv preprint arXiv:1804.03599 (2018).

Limitations of β-VAE

- Let us look at the VAE as a lossy compression
- For a RD-curve possible with the model structure,
- The VAE and the β-VAE only get one point on it

- This means one must train a new model according to
- Architecture, dataset, purpose, class, condition… etc.
- Sensitive hyperparameters tire us all the time

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

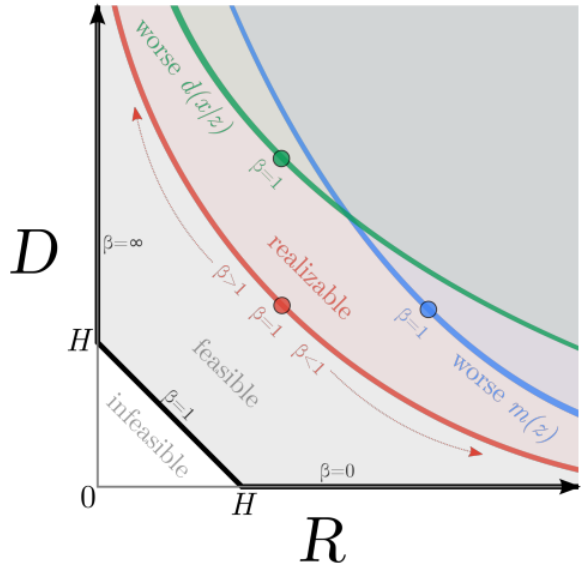


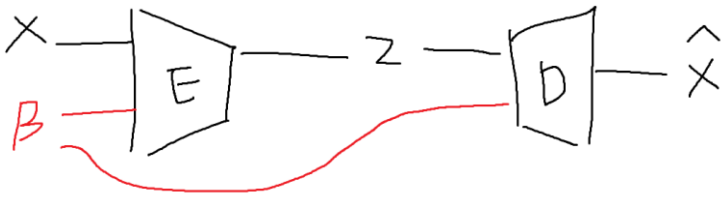
Figure 1. Schematic representation of the phase diagram in the *RD*-plane. The *distortion* (D) axis measures the reconstruction error of the samples in the training set. The *rate* (R) axis measures the relative KL divergence between the encoder and our own marginal approximation. The thick black lines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

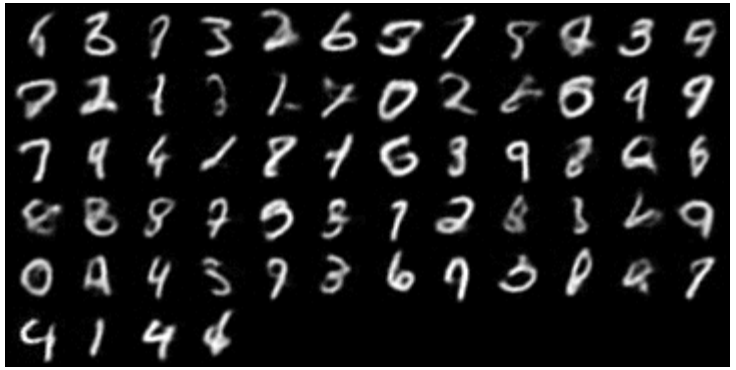
Alemi, Alexander, et al. "Fixing a broken ELBO." International conference on machine learning. PMLR, 2018.

Multi-β Model

- I believed that models with different βs would have strong associations
- So I implemented one VAE instance that can handle multiple βs
- This worked great but lacked novelty due to the existence of the prior work!

- I tried some analysis and proposed a novel heuristic, but it did not work well
- This is the study I did this summer





Rate-Distortion Curve

- This is the achieved RD-Curve, and the right side is from the references
- Why are these in log-like form?

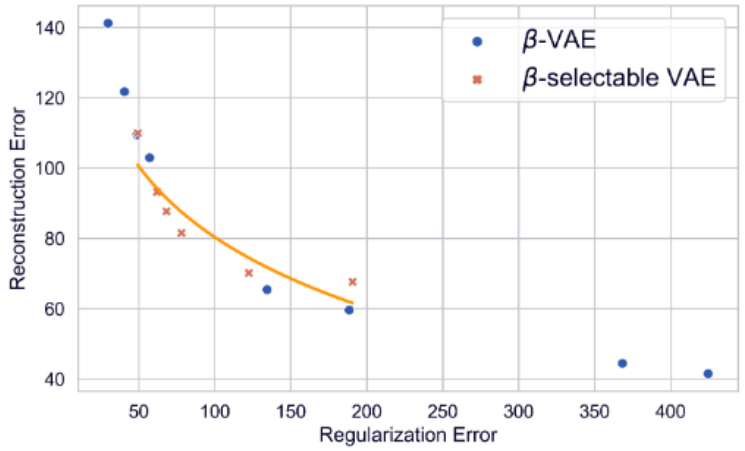
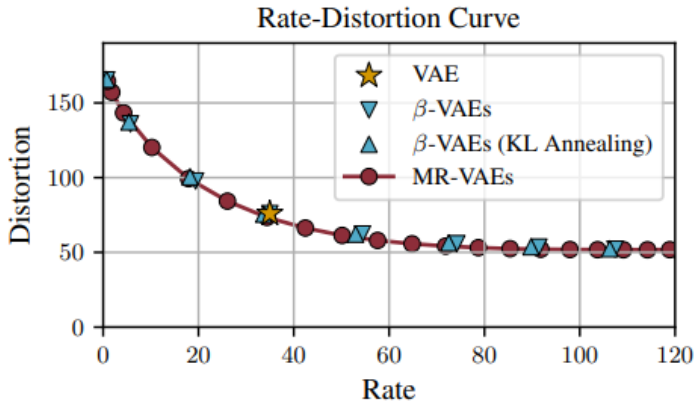


Figure 3: Reconstruction and regularization errors of ELBO for various βs. For ELBO, i.e. the learning objective, a single β-Selectable VAE is capable of approximate many β-VAEs. This graph can be interpreted as a rate-distortion curve.



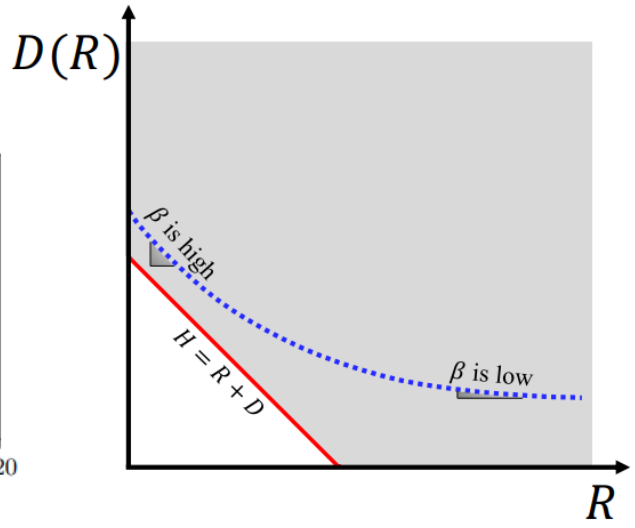


Figure 2: Schematic view of distortion-rate function. A red line corresponds to the theoretical lower bound of the rate and distortion. By varying β of β-VAE, we could achieve the points on a blue dashed curve, the sub-optimal distortion-rate function, which is best achievable with VAEs.

Bae, Juhan, et al. "Multi-Rate VAE: Train Once, Get the Full Rate-Distortion Curve." The Eleventh International Conference on Learning Representations. 2022. Seonho, George Adosoglou, and Panos M. Pardalos. "Interpreting rate-distortion of variational autoencoder and using model uncertainty for anomaly detection." Annals of Mathematics and Artificial Intelligence 90.7-9 (2022): 735-752.

Decoupling σ_x and β

- The β-VAE has been popular and there are numerous analyses on it
- What I doubt is how they dealt with the variance of the decoder σ_x
- In terms of formula, $\sigma_x(z)$ has a similar meaning to β
- The difference is that the σ_x itself remains the likelihood

- But the typical beta-VAE is pretty much the same

- Original VAE loss:

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

- Beta-VAE loss :

$$E_{z\sim q_\phi(z|x)}[MSE] + \beta D_{KL}(q_\phi(z|x)||p_\theta(z))$$

- Implemented correct ELBO :

$$E_{z\sim q_\phi(z|x)}\left[\frac{MSE}{2\sigma^2(z)} + \frac{1}{2}\log 2\pi\sigma^2(z)\right] + D_{KL}(q_\phi(z|x)||p_\theta(z))$$

Decoupling σ_x and β

- Many implementations keep σ_x constant and exclude it from optimization
- My claim is: studies so far have mixed the effects of σ_x and β
- Separating these explicitly will make the effect of each clearer
- This is the ***first research direction***

- But the typical beta-VAE is pretty much the same

- Original VAE loss:

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

- Beta-VAE loss :

$$E_{z\sim q_\phi(z|x)}[MSE] + \beta D_{KL}(q_\phi(z|x)||p_\theta(z))$$

- Implemented correct ELBO :

$$E_{z\sim q_\phi(z|x)}\left[\frac{MSE}{2\sigma^2(z)} + \frac{1}{2}\log 2\pi\sigma^2(z)\right] + D_{KL}(q_\phi(z|x)||p_\theta(z))$$

Decoupling σ_x and β

- Perhaps the analytically optimal $\sigma_x = \sqrt{MSE}$ works
- In this case, the reconstruction term becomes $\log MSE$
- The toy example had an unusual result: it has linear RD-Curve…?

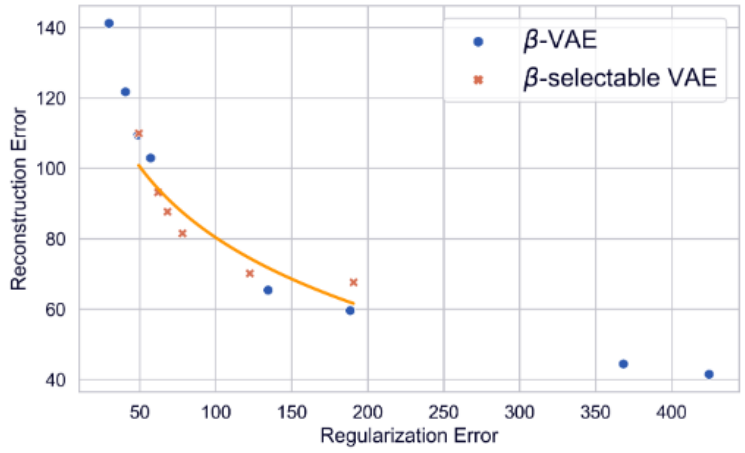


Figure 3: Reconstruction and regularization errors of ELBO for various βs. For ELBO, i.e. the learning objective, a single β-Selectable VAE is capable of approximate many β-VAEs. This graph can be interpreted as a rate-distortion curve.

- But the typical beta-VAE is pretty much the same
- Original VAE loss:

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

- Beta-VAE loss :

$$E_{z\sim q_\phi(z|x)}[MSE] + \beta D_{KL}(q_\phi(z|x)||p_\theta(z))$$

- Implemented correct ELBO :

$$E_{z\sim q_\phi(z|x)}\left[\frac{MSE}{2\sigma^2(z)} + \frac{1}{2}\log 2\pi\sigma^2(z)\right] + D_{KL}(q_\phi(z|x)||p_\theta(z))$$

Aggregated Posterior

- This means $q_\phi(z) = \frac{1}{N}\sum_x^N q_\phi(z|x)$
- Sampling would be good if $q_\phi(z) \approx p(z)$

- In general, as β increases,
- i.e. the regularization term decreases,
- The above property is satisfied

- But the shape of the posterior is different
- For the same Rate…?

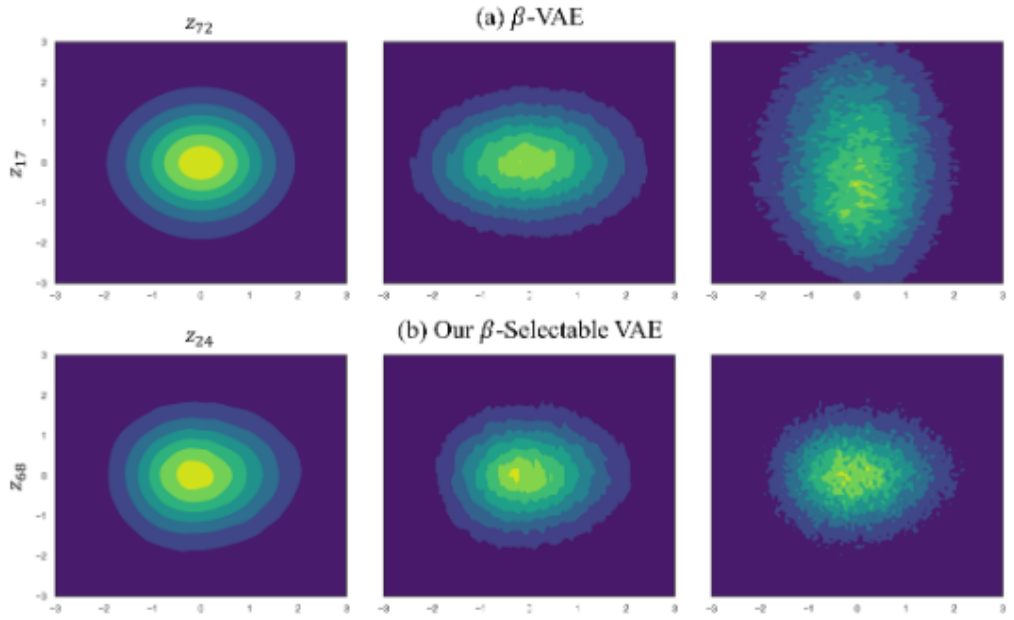


Figure 2: Aggregated posteriors for arbitrarily chosen latent channels. Each row shares a model and channels, and distributions are placed in order of decreasing β. The posteriors of our model (b) are *aligned* relative to their counterparts (a).

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

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

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

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

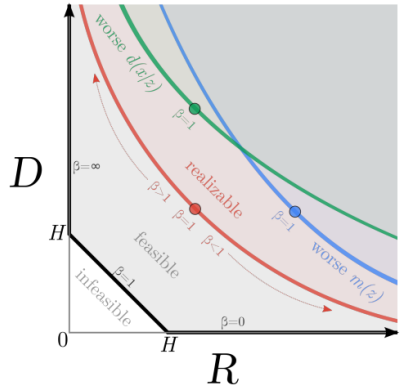


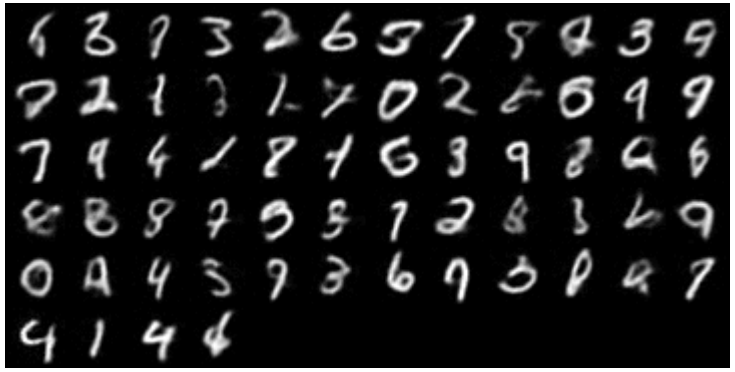
Figure 1. Schematic representation of the phase diagram in the *RD*-plane. The *distortion* (D) axis measures the reconstruction error of the samples in the training set. The *rate* (R) axis measures the relative KL divergence between the encoder and our own marginal approximation. The thick black lines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



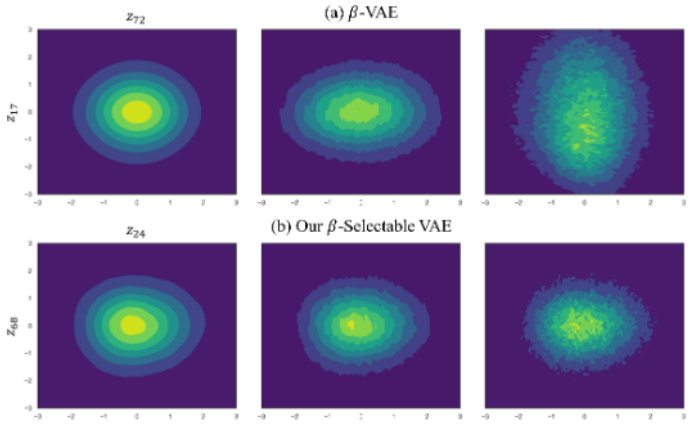


Figure 2: Aggregated posteriors for arbitrarily chosen latent channels. Each row shares a model and channels, and distributions are placed in order of decreasing β. The posteriors of our model (b) are *aligned* relative to their counterparts (a).

Thank you for listening

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