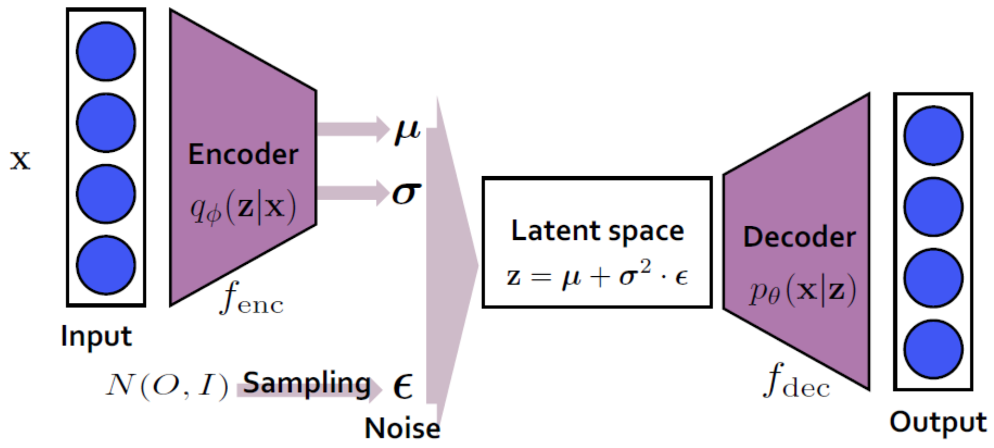
Isolating Beta from Sigma in Gaussian VAE

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

Background

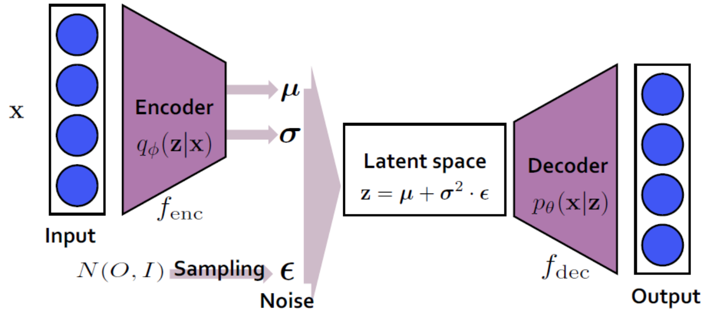
Variational Autoencoder

- VAE is a *Latent Variable Model*
- Statistically speaking, it infers latent Z from observable X
- X will be a dataset in the ML or DL field



Variational Autoencoder (contd.)

- Specifically, VAE employs variational inference
- We can model $p_\theta(X|Z)$, but then $p_\theta(Z|X)$ will generally be intractable
- So we train the model using its approximation $q_\phi(Z|X)$
- It will be a process of $X \rightarrow Z \rightarrow X$
- For special cases where $p_\theta(Z|X)$ is tractable, see 'Flow-based model'
- [1] Kingma, Diederik P., and Max Welling. "Auto-encoding variational bayes." arXiv preprint arXiv:1312.6114 (2013).



Variational Autoencoder (contd.)

- It learns the lower bound of the likelihood function as the objective
- $p_\theta(X) \geq E_{z \sim q_\phi(Z|X)}[\log p_\theta(X|Z)] - D_{KL}(q_\phi(Z|X)||p(Z))$ (ELBO)
- Red one is the reconstruction loss, the other is the regularization loss
- This equation is completely tractable with a few assumption

Where VAE can be used

- VAE has two main characters:

- First, it can be used as generative model
- We can sample $p_\theta(x)$ with $p_\theta(x|z)$,
- if we set $p(z)$ to be an easy-to-sample distribution

- Second, it produces compressed latent information
- Remember the $X \to Z \to X$, and generally we set $\dim Z \leq \dim X$

Where VAE can be used (contd.)

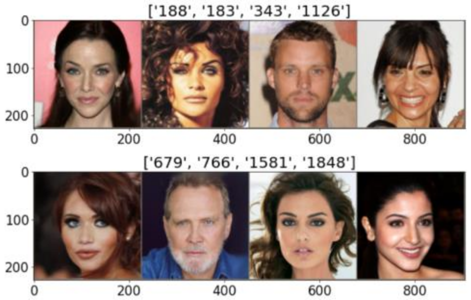
- Therefore, VAE can be used to obtain good samples
- Or to obtain lower-dimension representation
- e.g. Sentence generation[2], Image compressing[3], Outlier detection[4], …
- Molecular generation[5], Unsupervised learning (to get representation), etc.

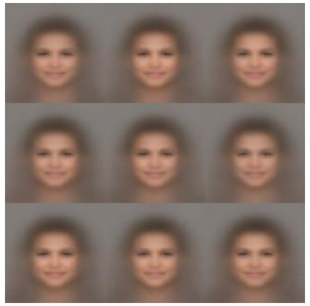
- [2] Bowman, Samuel R., et al. "Generating sentences from a continuous space." arXiv preprint arXiv:1511.06349 (2015).
- [3] Ballé, Johannes, et al. "Variational image compression with a scale hyperprior." arXiv preprint arXiv:1802.01436 (2018).
- [4] An, Jinwon, and Sungzoon Cho. "Variational autoencoder based anomaly detection using reconstruction probability." Special lecture on IE 2.1 (2015): 1-18.
- [5] Jin, Wengong, Regina Barzilay, and Tommi Jaakkola. "Junction tree variational autoencoder for molecular graph generation." International conference on machine learning. PMLR, 2018.

- These papers are highly cited examples, so read on if you are interested!

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)



beta-VAE

- β-VAE is the most famous improvement of VAE
- β-VAE: $-E_{z\sim q_{\phi}(Z|X)}[\log p_{\theta}(X|Z)] + \boldsymbol{\beta} D_{KL}(q_{\phi}(Z|X)||p(Z))$
- This balances two losses; manage the trade-off between the two
- It is known to be able to adjust posterior collapse[6], blurry output[7, 8], poor sampling[7, 8], and latent disentanglement[7, 9]

- [6] Lucas, James, et al. "Understanding posterior collapse in generative latent variable models." (2019).
- [7] Higgins, Irina, et al. "beta-vae: Learning basic visual concepts with a constrained variational framework." International conference on learning representations. 2016.
- [8] Alemi, Alexander, et al. "Fixing a broken ELBO." International conference on machine learning. PMLR, 2018.
- [9] Burgess, Christopher P., et al. "Understanding disentangling in $\beta $-VAE." arXiv preprint arXiv:1804.03599 (2018).

Rate-Distortion Curve

- $X \to Z \to X$ also looks like compression and decompression
- We can apply the rate-distortion curve used in information theory

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

- The red is the reconstruction loss, so it means Distortion
- The blue is the regularization loss, so it means Rate
- β-VAE is expressed with these two values[8]

Rate-Distortion Curve (contd.)

- So the β-VAE is a point on the valid RD curve
- And the β is the parameter that causes it to move along it

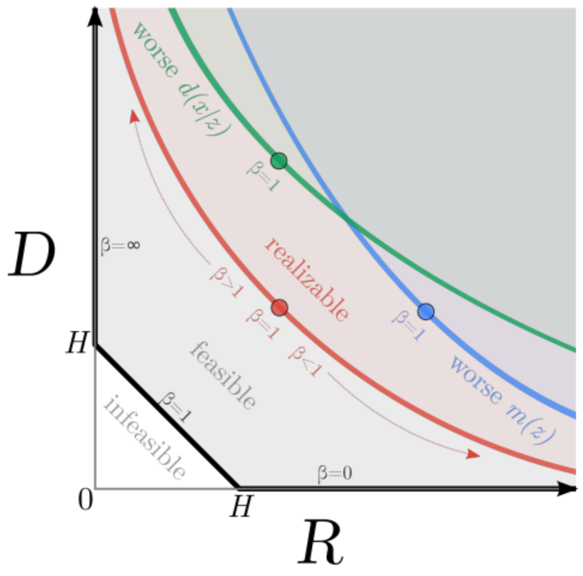


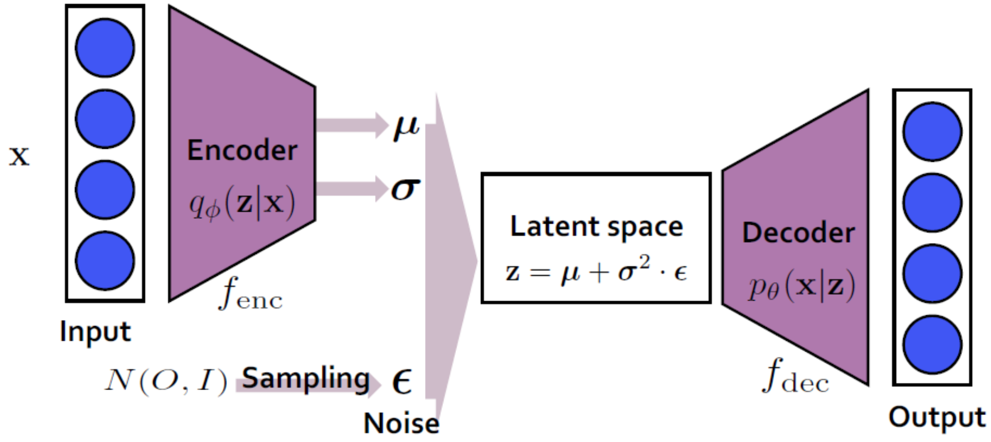
Figure 1. Schematic representation of the phase diagram in the RD-plane. The *distortion* (D) axis measures the reconstruction error of the samples in the training set. The *rate* (R) axis measures the relative KL divergence between the encoder and our own marginal approximation. The thick black lines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

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

Claim

Implementation of VAE

- $p_\theta(X|Z)$ and $q_\phi(Z|X)$ are often modeled as Gaussian
- $p_\theta(X|Z) \sim N(\mu_X(Z), \sigma_X(Z)I)$ – the shared diagonal covariance
- $q_\phi(Z|X) \sim N(\mu_Z(X), \sigma_Z(X))$ – the diagonal covariance
- Its diagonal covariance is known as an important assumption; see [9, 10]
- [10] Kumar, Abhishek, and Ben Poole. "On Implicit Regularization in $ β $-VAEs." International Conference on Machine Learning. PMLR, 2020.



Implementation of VAE (contd.)

- $\sigma_X(Z)$ is usually set to be a *constant*
- Perhaps because learning $\sigma_X(Z)$ introduces instability
- $\sigma_X(Z)$ sometimes goes to 0 and this makes an infinite gradient

Greens can be infinitely large or small

Reconstruction loss (Distortion)

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

Regularization loss (Rate)

Connection between sigma and beta

- Look at the formula carefully…

$$Loss = -E[\log p_\theta(X|Z)] + D_{KL}(q_\phi(Z|X)||p(Z))$$
$$= -E\left[\frac{\left(X-\mu_X(Z)\right)^2}{2\sigma_X^2} + \frac{\log 2\pi\sigma_X^2}{2}\right] + D_{KL}(q_\phi(Z|X)||p(Z))$$

- So if we set σ_X as a constant,

$$2\sigma_X^2 Loss = -E\left[\left(X-\mu_X(Z)\right)^2\right] + 2\sigma_X^2 D_{KL}(q_\phi(Z|X)||p(Z)) + C$$
$$= -E\left[\left(X-\mu_X(Z)\right)^2\right] + \beta D_{KL}(q_\phi(Z|X)\big|\big|p(Z)\big) + C$$

- It becomes β-VAE objective

Connection between sigma and beta (contd.)

- This is a pretty interesting perspective
- Previous studies have focused on this aspect

- But, σ_X and β are definitely different!
- This has been pointed out before: see [11]

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

Learnable sigma

- The two objectives become the same when σ_X is set as a constant
- It would be different if it were a learnable σ_X!
- The log-sigma term can no longer be the constant C

$$-E\left[\frac{\left(X-\mu_X(Z)\right)^2}{2\sigma_X^2}+\frac{\log 2\pi\sigma_X^2}{2}\right]+D_{KL}(q_\phi(Z|X)||p(Z))$$

$$-E\left[\left(X-\mu_X(Z)\right)^2\right]+\beta D_{KL}\left(q_\phi(Z|X)\middle\|p(Z)\right)+C$$

Learnable sigma (contd.)

- Where does the log-sigma term come from?

$$Loss = -E[\log p_\theta(X|Z)] + D_{KL}(q_\phi(Z|X)||p(Z))$$
$$= -E\left[\frac{\left(X-\mu_X(Z)\right)^2}{2\sigma_X^2} + \frac{\log 2\pi\sigma_X^2}{2}\right] + D_{KL}(q_\phi(Z|X)||p(Z))$$

- It comes from the normalizer of the Gaussian pdf

Support	$x \in \mathbb{R}$
PDF	$\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$

- Intuitively, leaving the normalizer constant or ignoring it
- Would lead to pathological prior knowledge

Learnable sigma is integral

- From a theoretical perspective, learnable σ_X turns out to be important
- See [12, 13, 14]

- [12] Dai, Bin, and David Wipf. "Diagnosing and enhancing VAE models." arXiv preprint arXiv:1903.05789 (2019).
- [13] 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.
- [14] Koehler, Frederic, et al. "Variational autoencoders in the presence of low-dimensional data: landscape and implicit bias." arXiv preprint arXiv:2112.06868 (2021).

Implementation of Learnable sigma

- There are already some practical studies on learnable σ_X [15, 16]
- These use some novel ideas to reliably introduce σ_X into learning
- But *(even though these are studies of learnable one)* they emphasize that it is related to the β [12, 15]
- Those such as [15] have very good results, but they simplify their work to finding the optimal β

- [15] Rybkin, Oleh, Kostas Daniilidis, and Sergey Levine. "Simple and effective VAE training with calibrated decoders." International Conference on Machine Learning. PMLR, 2021.
- [16] Takahashi, Hiroshi, et al. "Student-t Variational Autoencoder for Robust Density Estimation." IJCAI. 2018.

Isolating beta from sigma

- I believe that the β and the σ_X are different
- …when it comes to learnable σ_X

- I put both together and show the situation that is better than using one
- It is tested on several popular computer vision datasets
- This means that there are situations where β and σ_X are different
- And good when used correctly

Experiment

Experiment 1. Rate-Distortion Curve

- The design purpose of β can be clarify with RD curve
- Let us look at two commonly used assumption: *(need references!)*

Let the $Loss = -E\left[\left(X - \mu_X(Z)\right)^2\right] + KD_{KL}(q_\phi(Z|X)\left|\right|p(Z)) + C$

1. $\sigma_X = \frac{1}{2}$ and $\beta = K$ – the sigma is a constant and the beta is the beta

2. $\sigma_X = \frac{K}{2}$ and $\beta = 1$ or something – the $\beta = 2\sigma_X$

Experiment 1. Rate-Distortion Curve (contd.)

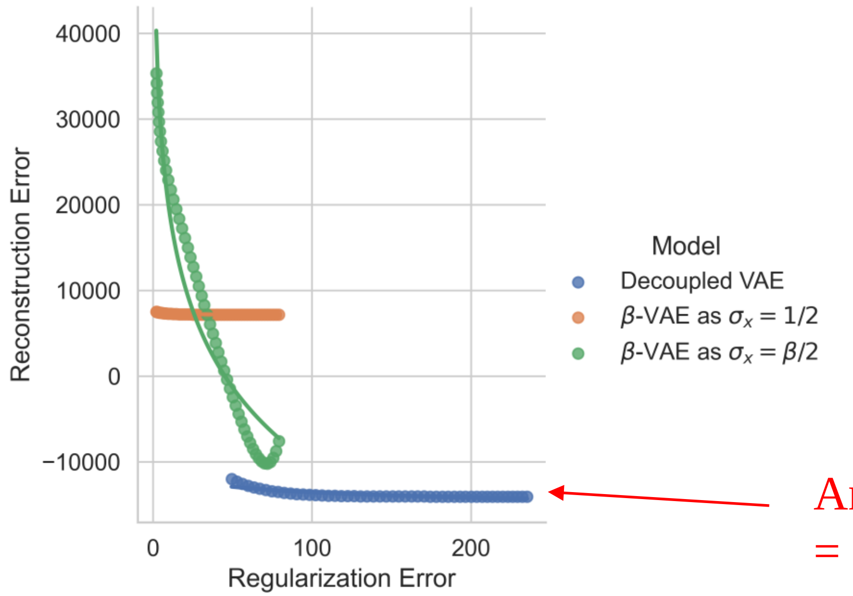
1. $\sigma_X = \frac{1}{2}$ and $\beta = K$ – the sigma is a constant and the beta is the beta
2. $\sigma_X = \frac{K}{2}$ and $\beta = K$ or something – the $\beta = 2\sigma_X$

$$-E\left[\frac{\left(X - \mu_X(Z)\right)^2}{2\sigma_X^2} + \frac{\log 2\pi\sigma_X^2}{2}\right] + D_{KL}(q_\phi(Z|X)||p(Z))$$

- Can you see that the RD value changes in the two cases?
- Even though the model is the same,
- it changes depending on how you look at it

Experiment 1. Rate-Distortion Curve (contd.)

- I plan to show that applying β and σ_X separately (decoupled one) is better in terms of RD than in both of previous cases
- This is the result of a rough experiment



Anyway, it is further down
= It performs better

Experiment 2. Proxy Metric

- In the end, the evaluation of the generative model is based on metrics
- I will show that the best decoupled model is better than the best baseline models through proxy metrics e.g. FID score

	log	β	CelebA	MNIST
β-VAE	X	100.0	198.64	
β-VAE	X	10.0	112.07	344.15
β-VAE	X	1.0	70.62	100.86
β-VAE	X	0.1	94.22	79.54
β-VAE	X	0.01	86.86	124.82
β-VAE	X	0.001	266.36	
β-VAE	O	100.0	72.49	
β-VAE	O	10.0	58.82	32.38
β-VAE	O	1.0	74.26	42.99
β-VAE	O	0.1	335.55	67.35
β-VAE	O	0.01	69.05	63.65
β-VAE	O	0.001	235.20	

Lower is better
Remarkable difference…

Thank you

- This is (probably) the final refined version of an argument
- … which I have been making for months

- I am planning to write a paper based on this development
- And always thirsty for better mathematical proofs or ingenious experiments
- If you have any idea, please discuss it any time

- Any question?

References

- [1] Kingma, Diederik P., and Max Welling. "Auto-encoding variational bayes." arXiv preprint arXiv:1312.6114 (2013).
- [2] Bowman, Samuel R., et al. "Generating sentences from a continuous space." arXiv preprint arXiv:1511.06349 (2015).
- [3] Ballé, Johannes, et al. "Variational image compression with a scale hyperprior." arXiv preprint arXiv:1802.01436 (2018).
- [4] An, Jinwon, and Sungzoon Cho. "Variational autoencoder based anomaly detection using reconstruction probability." Special lecture on IE 2.1 (2015): 1-18.
- [5] Jin, Wengong, Regina Barzilay, and Tommi Jaakkola. "Junction tree variational autoencoder for molecular graph generation." International conference on machine learning. PMLR, 2018.
- [6] Lucas, James, et al. "Understanding posterior collapse in generative latent variable models." (2019).
- [7] Higgins, Irina, et al. "beta-vae: Learning basic visual concepts with a constrained variational framework." International conference on learning representations. 2016.
- [8] Alemi, Alexander, et al. "Fixing a broken ELBO." International conference on machine learning. PMLR, 2018.
- [9] Burgess, Christopher P., et al. "Understanding disentangling in β-VAE." arXiv preprint arXiv:1804.03599 (2018).
- [10] Kumar, Abhishek, and Ben Poole. "On Implicit Regularization in $ β $-VAEs." International Conference on Machine Learning. PMLR, 2020.
- [11] Lucas, James, et al. "Don't blame the elbo! a linear vae perspective on posterior collapse." Advances in Neural Information Processing Systems 32 (2019).
- [12] Dai, Bin, and David Wipf. "Diagnosing and enhancing VAE models." arXiv preprint arXiv:1903.05789 (2019).
- [13] 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.
- [14] Koehler, Frederic, et al. "Variational autoencoders in the presence of low-dimensional data: landscape and implicit bias." arXiv preprint arXiv:2112.06868 (2021).
- [15] Rybkin, Oleh, Kostas Daniilidis, and Sergey Levine. "Simple and effective VAE training with calibrated decoders." International Conference on Machine Learning. PMLR, 2021.
- [16] Takahashi, Hiroshi, et al. "Student-t Variational Autoencoder for Robust Density Estimation." IJCAI. 2018.

Thank you for listening

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