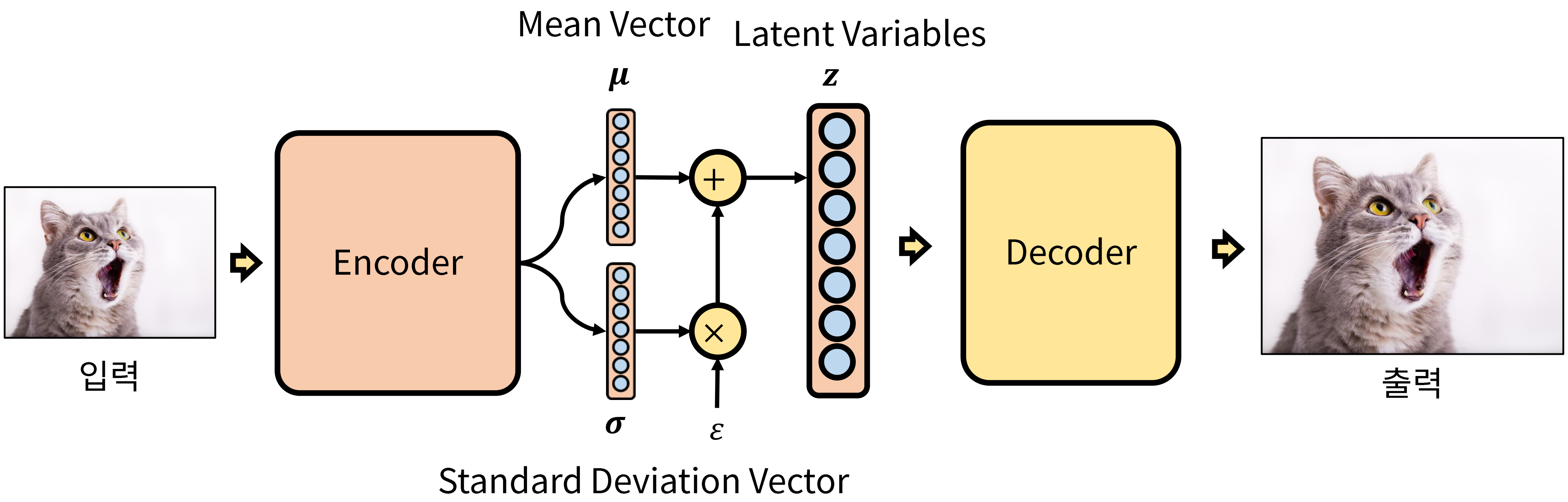
Chapter 06. 무엇이든 진짜처럼 생성하는 생성 모델(Generative Networks)

Variational Autoencoder

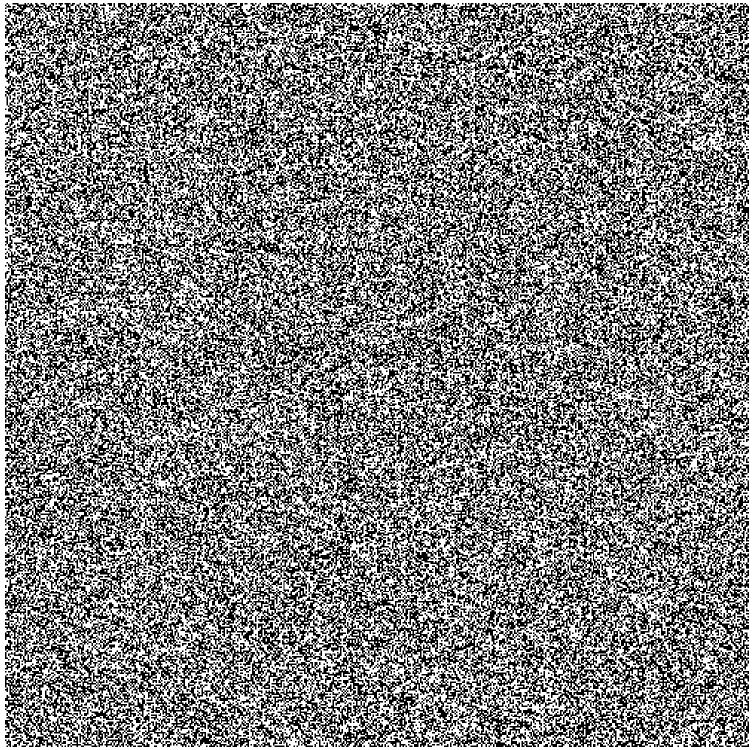
VAE Variational Autoencoder



GAN에 대해서 먼저 알아보았기 때문에, VAE는 조금 더 쉽게 이해할 수 있을 것이다!

Why VAE?

$P(X = x_1)$



$P(X = x_2)$

$$P(X = x_1) = P(X = x_2)$$

임의로 영상을 생성할 경우, 좌측과 우측 영상이 발생할 확률은 동일하다.

Natural Image는 전체 영상 Domain 중에서 매우 Sparse하다.

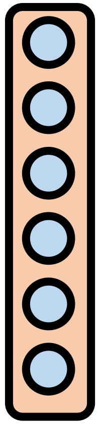
Manifold Learning

$\boldsymbol{x} \in \mathbb{R}^N$

Manifold Learning

$\boldsymbol{z} = f(\boldsymbol{x}|\boldsymbol{\theta})$

Embedding



$\boldsymbol{z} \in \mathbb{R}^M \ (N > M)$

더 낮은 차원으로$(N > M)$ 변환하는 것을 Embedding이라 하고,

이 Embedding Function을 학습하는 것을 Manifold Learning이라 한다.

Manifold Learning Example

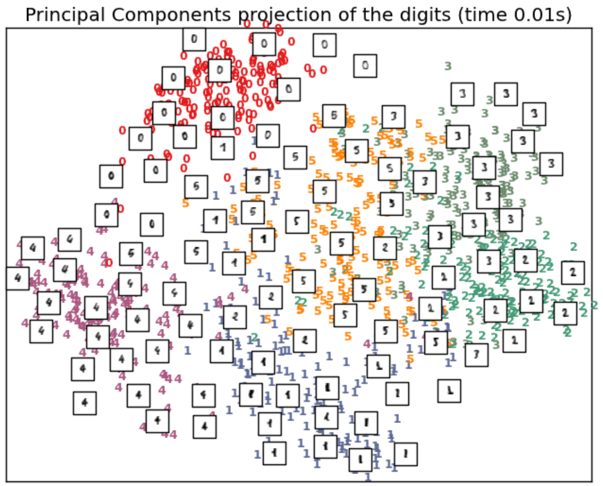
Image Space

$X \in \mathbb{R}^{8\times 8}$

$\boldsymbol{z} = f(X|\boldsymbol{\theta})$

Embedding

Latent Space

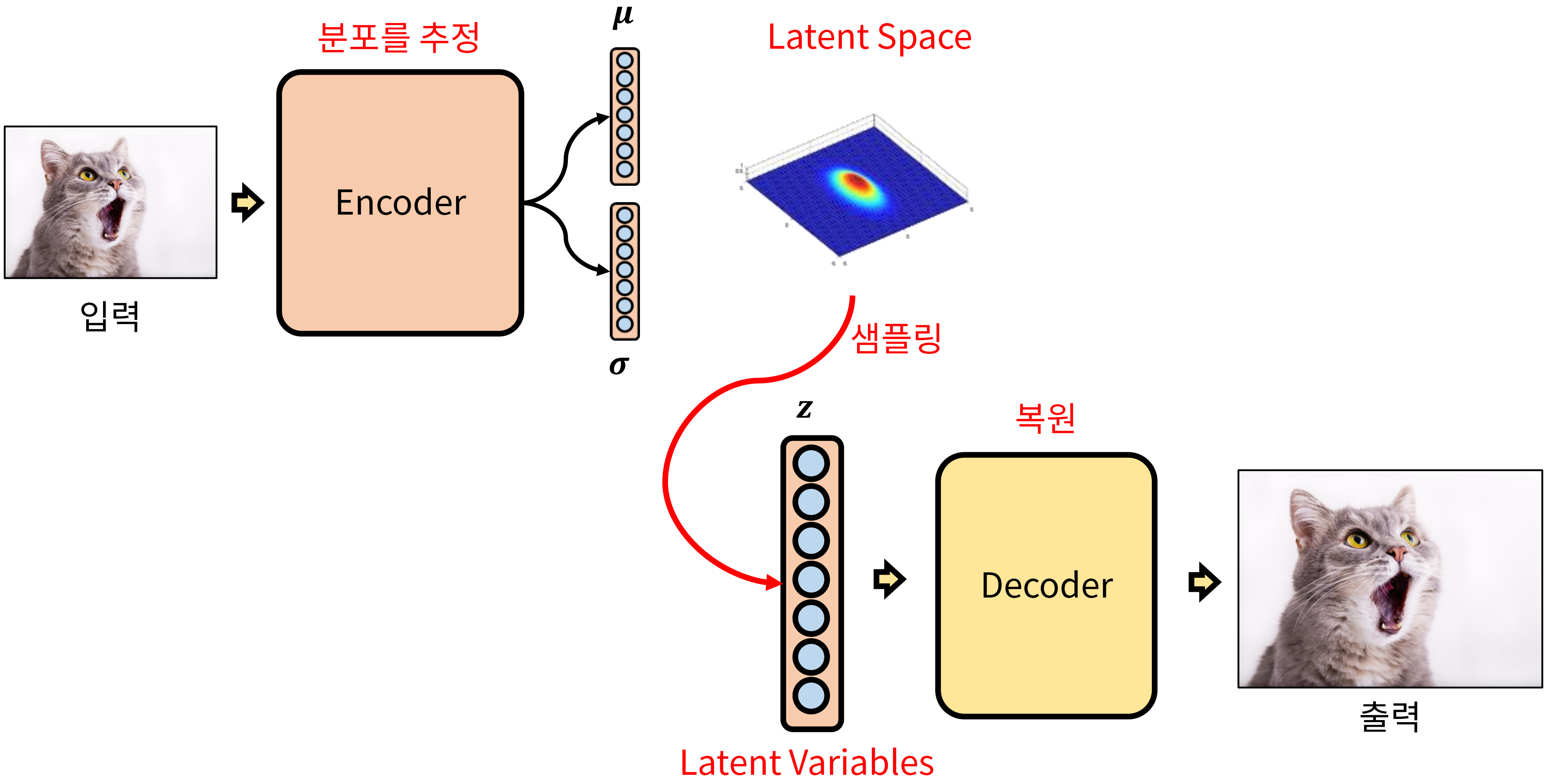


$\boldsymbol{z} \in \mathbb{R}^2$

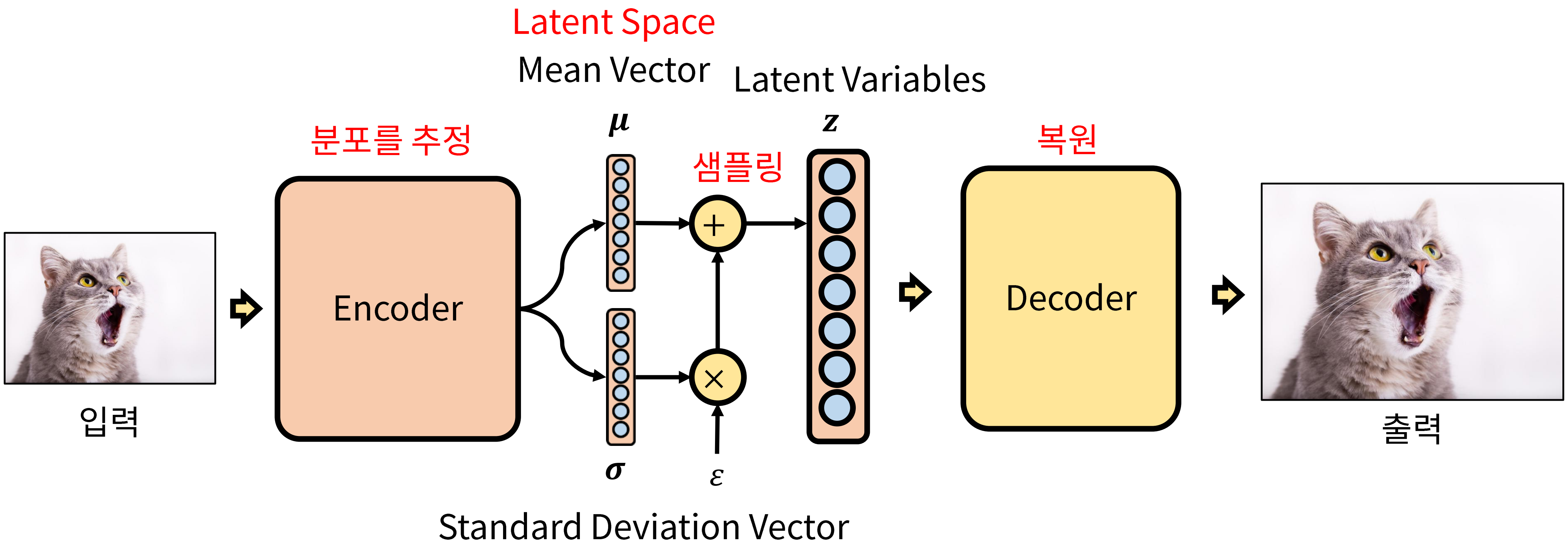
64차원 → 2차원의 Manifold Learning의 예 (PCA)

Image Space에서 생성하는 것 보다 Latent Space에서 생성하기가 훨씬 쉽다.

VAE Structure (1/2)



VAE Structure (2/2)



앞의 스토리를 보고 나니 좀 더 느낌이 오지 않는가?

KL Divergence Kullback-Leibler Divergence

$$D_{KL}(p||q) = \int p(x)\log\frac{p(x)}{q(x)}dx$$

$$= \int p(x)\log p(x)\,dx - \int p(x)\log q(x)\,dx$$

$$\neq D_{KL}\ (q||p)$$

KL Divergence는 두 확률 분포가 다른 정도를 나타내는 것이 목적이다.

‘거리(Distance)’라고 부를 수는 없는데, 교환법칙이 성립하지 않기 때문이다.

KLD of Gaussian Distribution

$$N(\mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$$p(x) = N(\mu_1, \sigma_1)$$

$$q(x) = N(\mu_2, \sigma_2)$$

$$D_{KL}(p||q) = -\int p(x) \log q(x)\, dx + \int p(x) \log p(x)\, dx$$

$$= \frac{1}{2}\log(2\pi\sigma_2^2) + \frac{\sigma_1^2 + (\mu_1 - \mu_2)^2}{2\sigma_2^2} - \frac{1}{2}(1 + \log 2\pi\sigma_1^2)$$

$$= \boxed{\log\frac{\sigma_2}{\sigma_1} + \frac{\sigma_1^2 + (\mu_1 - \mu_2)^2}{2\sigma_2^2}} - \frac{1}{2}$$

두 Gaussian 분포의 비교는 KL Divergence로 쉽게 표현할 수 있다.

Evidence Lower Bound (ELBO) (1/3)

$$\log P(x_i) = \log \frac{P(x_i|z)p(z)}{P(z|x_i)}$$

$$= \underline{\log P(x_i|z) + P(z) - \log P(z|x_i) - \log P(z|x_i)}$$

Log-Likelihood를 최대화 하고 싶다.

(MLE)

$$\log P(x_i) = \log P(x_i) \int q(z|x_i) dz$$

$$= \int q(z|x_i) \underline{\log P(x_i)}\, dz$$

위 수식을 대입하고

하나씩 전개해 보자.

Evidence Lower Bound (ELBO) (2/3)

$$\log P(x_i) = \boxed{\int q(z|x_i)[\log P(x_i|z)} + \log P(z) - \log P(z|x_i)]dz$$

$$= \boxed{E_{q(z|x_i)}[\log P(x_i|z)]} + \int q(z|x_i)\log p(z)\,dz - \int q(z|x_i)\log P(z|x_i)\,dz \boxed{\pm \int q(z|x_i)\log q(z|x_i)\,dz}$$

$$= E_{q(z|x_i)}[\log P(x_i|z)] - \boxed{\int q(z|x_i)\log q(z|x_i)\,dz + \int q(z|x_i)\log p(z)\,dz}$$

$$+ \boxed{\int q(z|x_i)\log q(z|x_i)\,dz - \int q(z|x_i)\log P(z|x_i)\,dz}$$

$$= E_{q(z|x_i)}[\log P(x_i|z)] - \boxed{D_{KL}\big(q(z|x_i)||P(z)\big)} + \boxed{D_{KL}\big(q(z|x_i)||P(z|x_i)\big)}$$

Decoder의 사후확률은 알기 어렵다.

$$\geq E_{q(z|x_i)}[\log P(x_i|z)] - D_{KL}\big(q(z|x_i)||P(z)\big)$$

Evidence Lower Bound (ELBO) (3/3)

Loss Function

$$\log P(x_i) \geq \boxed{\underbrace{E_{q(z|x_i)}[\log P(x_i|z)]}_{\text{Reconstruction Error}} - \underbrace{D_{KL}\big(q(z|x_i)||P(z)\big)}_{\text{Regularization}}}$$

둘 모두 Gaussian 분포를 따른다.

$$\begin{cases} z \sim N(0,1) \\ q(z|x_i) \sim N\big(\mu_{q(x_i)}, \sigma_{q(x_i)}\big) \end{cases}$$

$$D_{KL}\big(q(z|x_i)||P(z)\big) = D_{KL}\left(N\big(\mu_{q(x_i)}, \sigma^2_{q(x_i)}\big)||N(0,1)\right)$$

$$= \sum_i -\log \sigma_{q(x_i)} + \frac{1}{2}\left(\sigma^2_{q(x_i)} + \mu^2_{q(x_i)} - 1\right)$$

Interesting Results

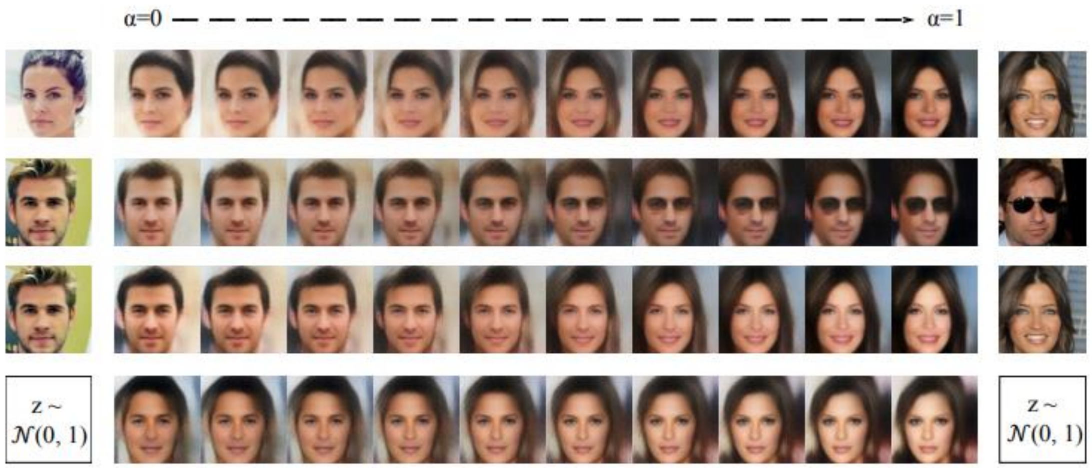


Figure 5. Linear interpolation for latent vector. Each row is the interpolation from left latent vector z_{left} to right latent vector z_{right}. e.g. $(1-\alpha)z_{left} + \alpha z_{right}$. The first row is the transition from a non-smiling woman to a smiling woman, the second row is the transition from a man without eyeglass to a man with eyeglass, the third row is the transition from a man to a woman, and the last row is the transition between two fake faces decoded from $z \sim \mathcal{N}(0, 1)$.

https://arxiv.org/pdf/1610.00291.pdf

Interesting Results

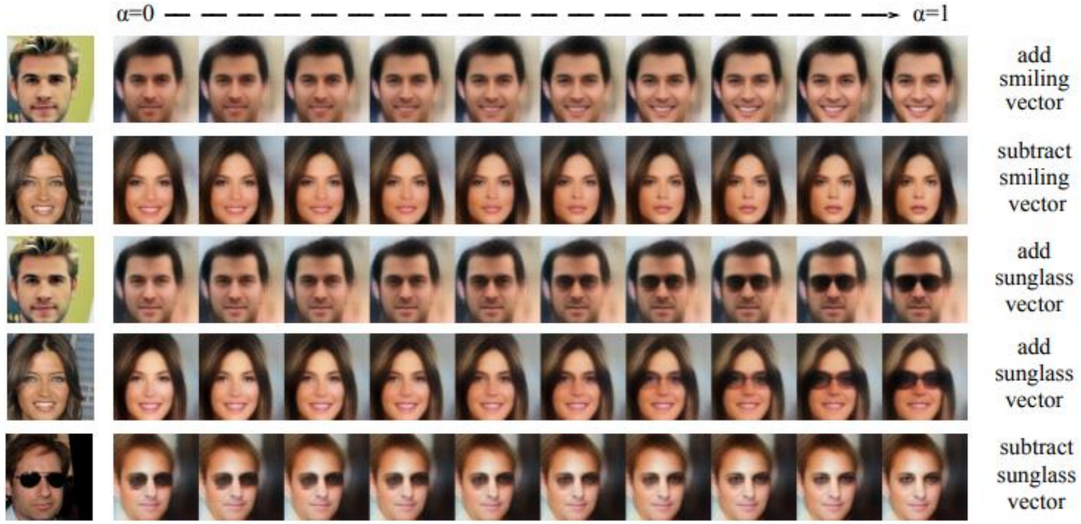


Figure 6. Vector arithmetic for visual attributes. Each row is the generated faces from latent vector z_{left} by adding or subtracting an attribute-specific vector, i.e., z_{left} + α $z_{smiling}$, where $\alpha = 0, 0.1, \ldots, 1$. The first row is the transition by adding a smiling vector with a linear factor α from left to right, the second row is the transition by subtracting a smiling vector, the third and fourth row are the results by adding a eyeglass vector to the latent representation for a man and women, and the last row shows results by subtracting an eyeglass vector.

https://arxiv.org/pdf/1610.00291.pdf