AUTOENCODER

정인호

Bayes rule?

Likelihood
How probable is the evidence given that our hypothesis is true?

Prior
How probable was our hypothesis *before* observing the evidence?

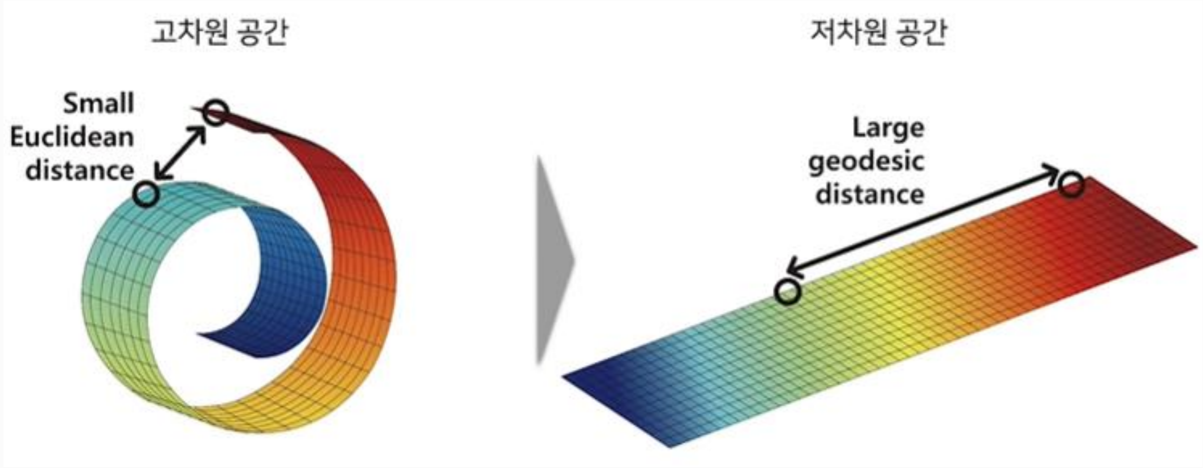
$$P(H \mid e) = \frac{P(e \mid H)\, P(H)}{P(e)}$$

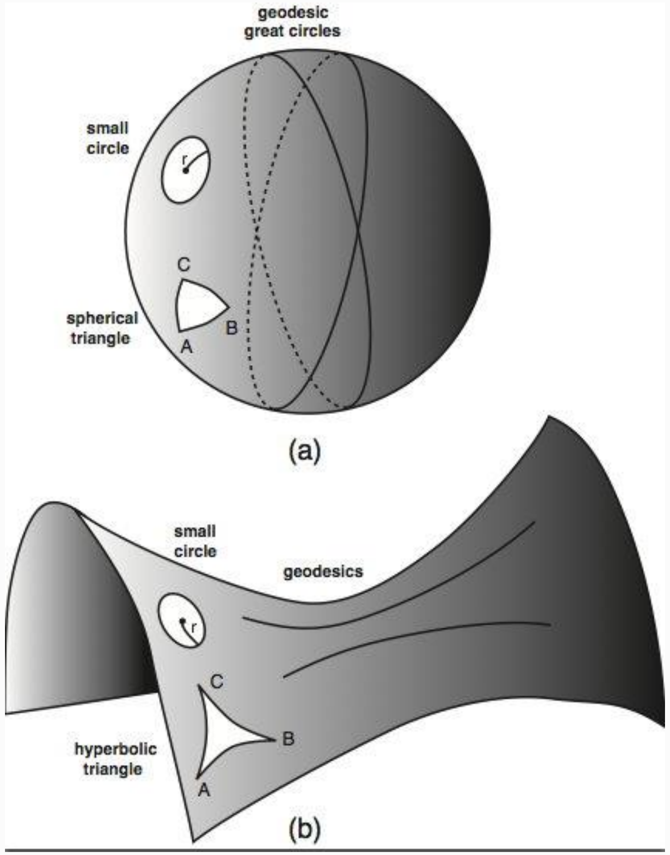
Posterior
How probable is our hypothesis given the observed evidence?
(Not directly computable)

Marginal
How probable is the new evidence under all possible hypotheses?
$P(e) = \sum P(e \mid H_i)\, P(H_i)$

Manifold?

In mathematics, a manifold is a topological space that locally resembles Euclidean space near each point.





KL-divergence

<u>정의</u> (쿨백-라이블러 발산, Kullback-Leibler divergence)

두 확률밀도함수 $p(x)$, $q(x)$에 대해서

$$KL(P \| q) := \int_{\mathbb{R}} p(x) \log\left(\frac{p(x)}{q(x)}\right) dx$$

를 두 확률분포 간의 <u>KL-divergence or relative entropy</u>라고 부른다.

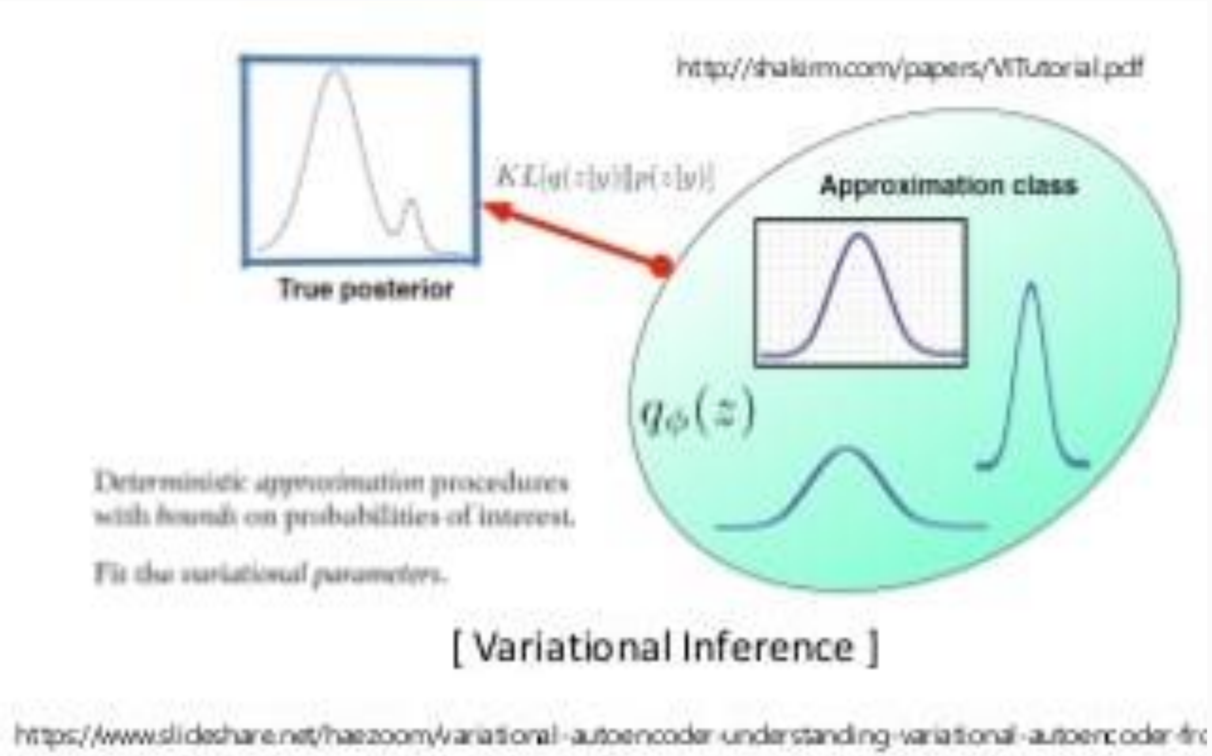
<u>KL-divergence의 성질</u>

- 임의의 p, q에 대해서 $KL(P \| q) \geq 0$
- $KL(P \| q) = 0$ if and only if $p = q$.

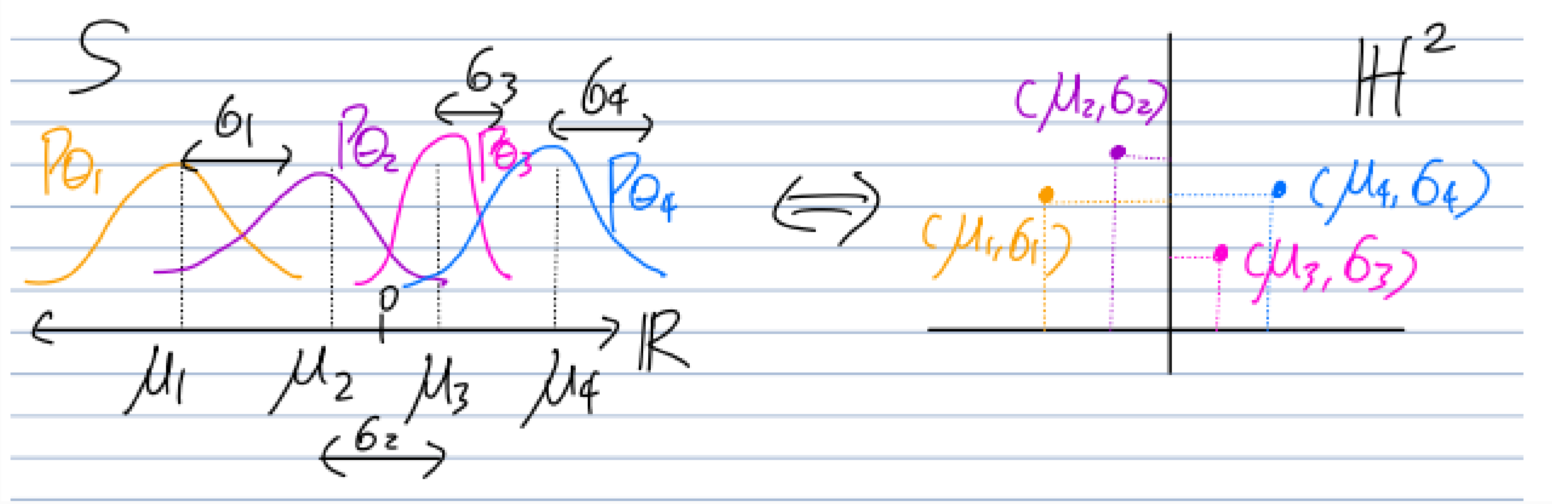
수학적으로 엄밀하게 말하면 metric은 아님. (삼각부등식 만족X)

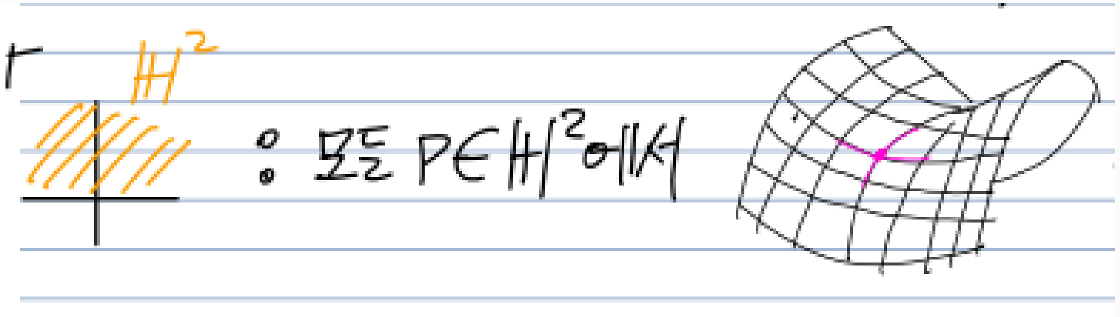
Information geometry

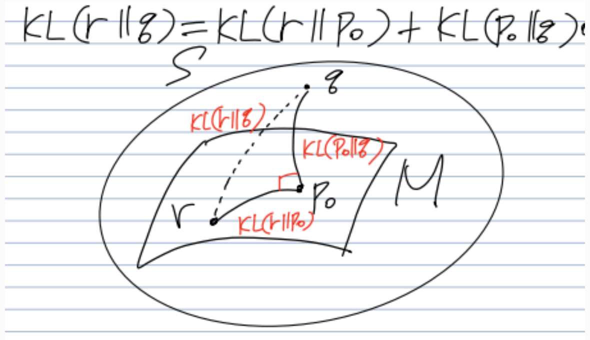
Q. The most reasonable Gaussian distribution?



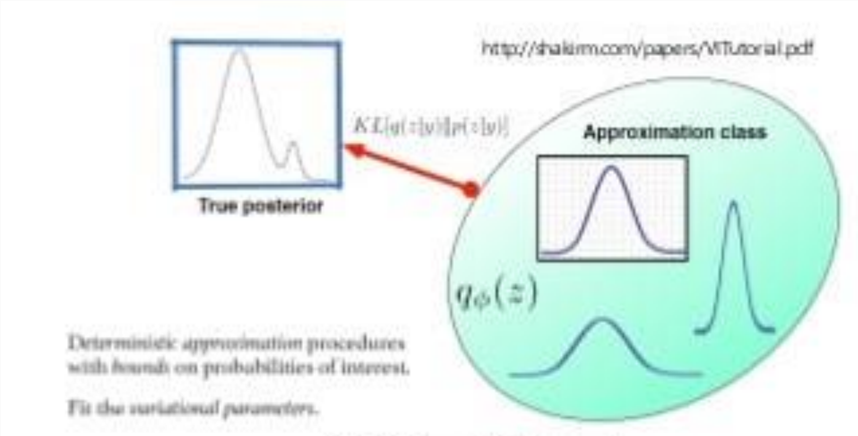
Information geometry





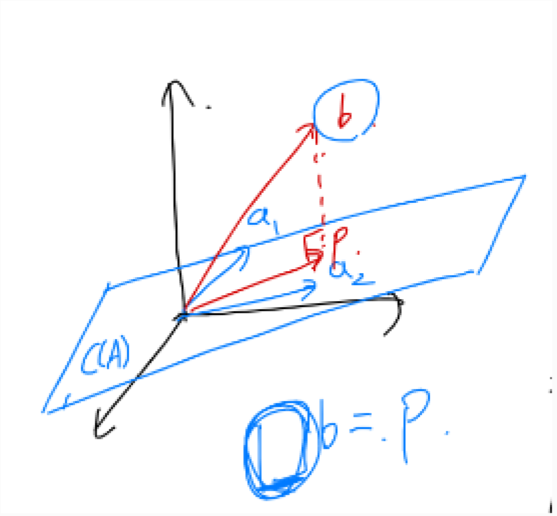


Information geometry

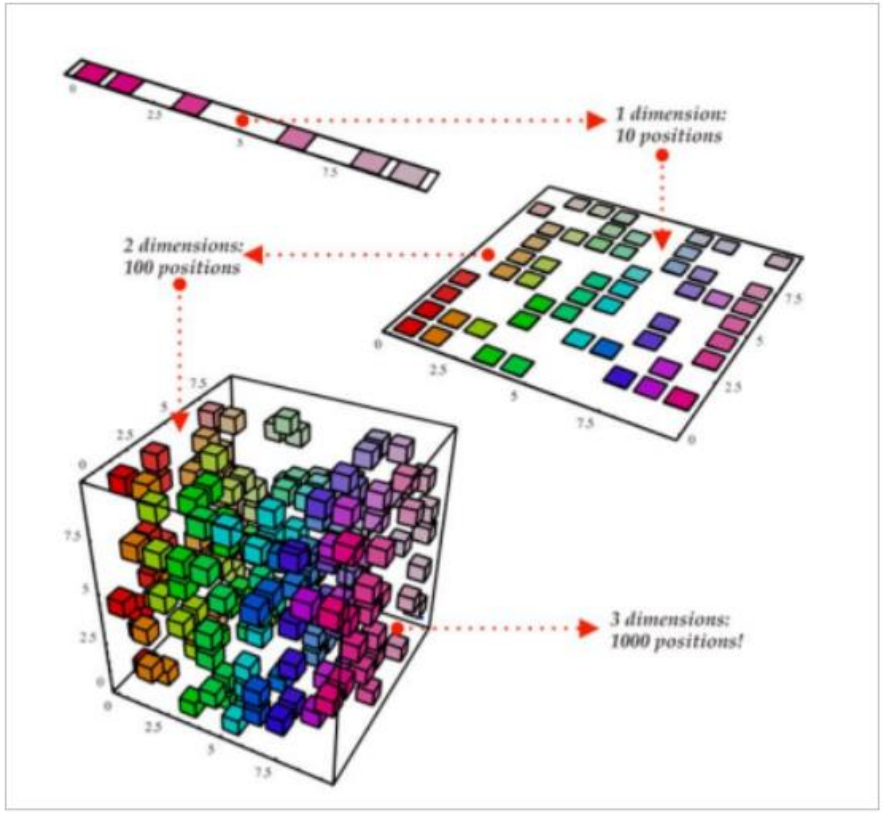


[Variational Inference]

https://www.slideshare.net/haezoom/variational-autoencoder-understanding-variational-autoencoder-fro



Curse of dimensionality

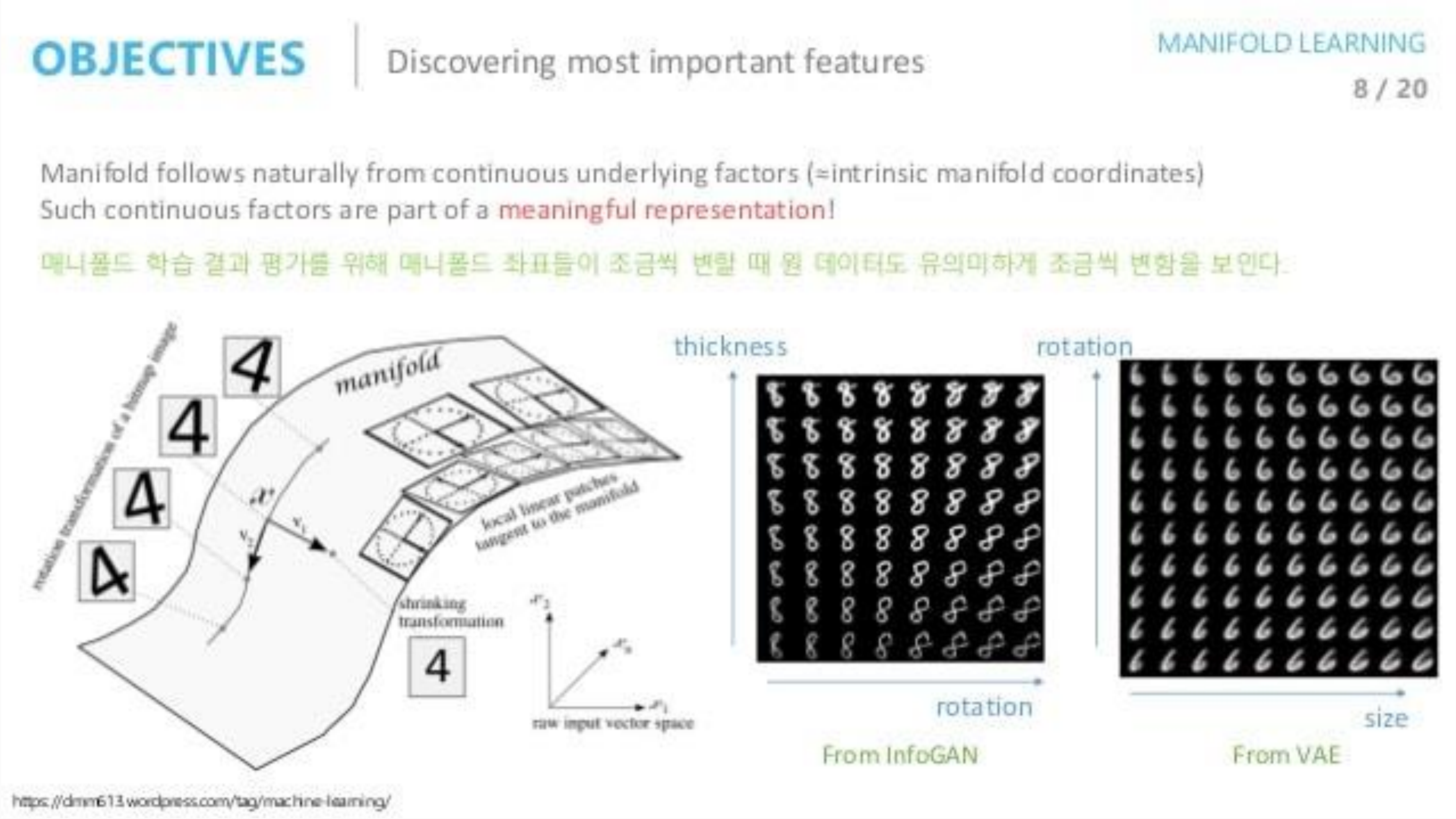


200*200 RGB image

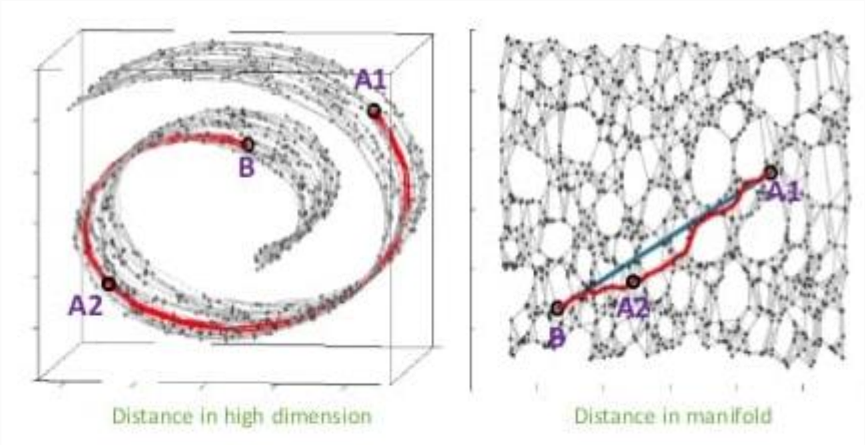
차원이 증가할수록 데이터의 분포 분석 또는 모델추정에 필요한 샘플 게이터의 개수가 기하급수적으로 증가

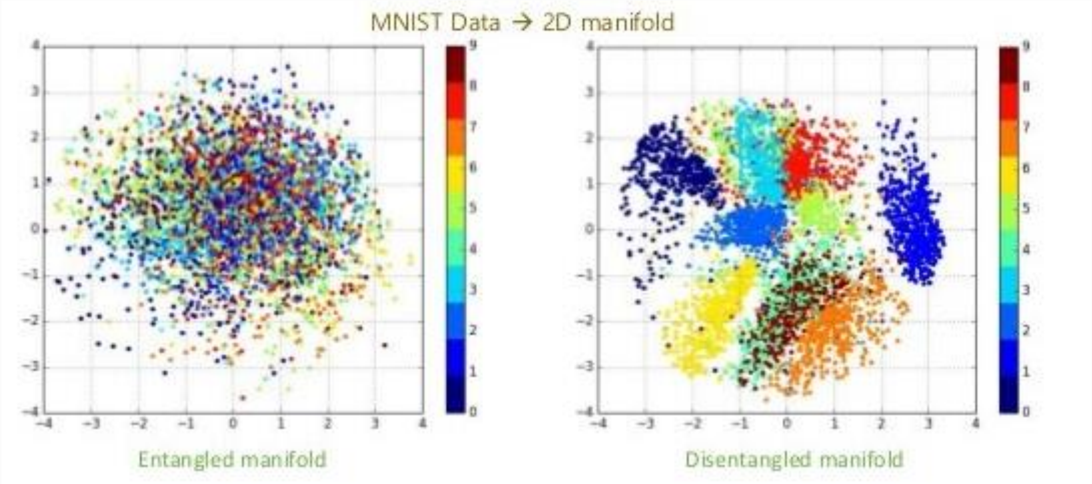
Manifold Hypothesis

Natural data in high dimensional spaces concentrates close to lower dimensional manifolds.



Manifold Hypothesis

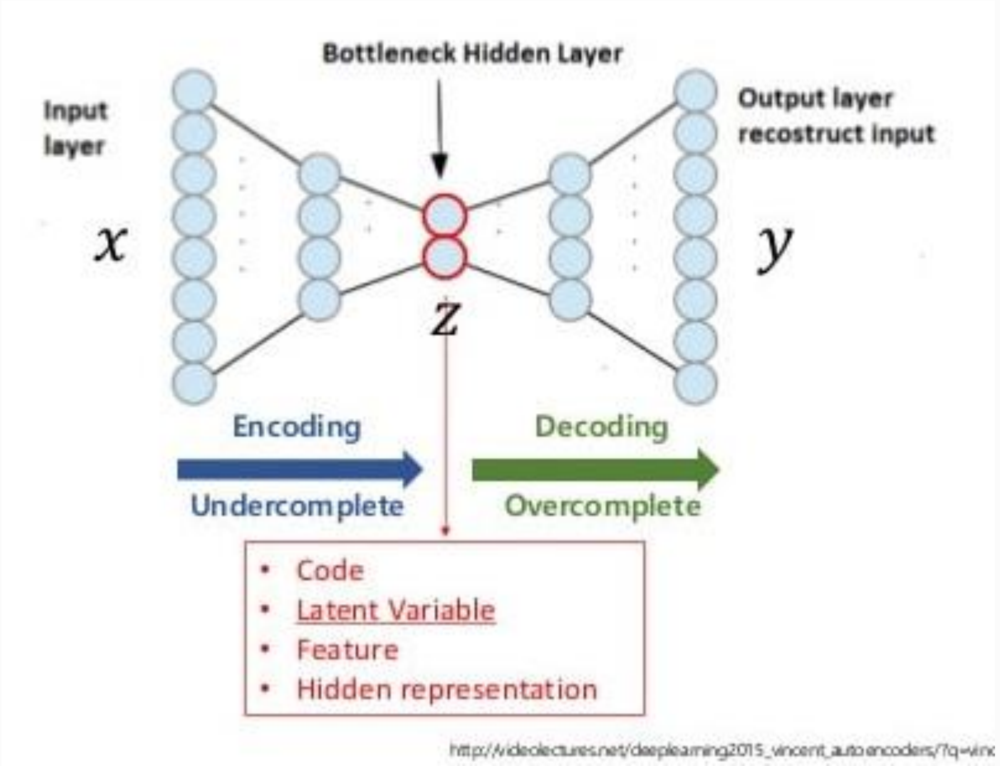




Q. 4차원 이상에서 Disentangled 확인하는 방법?

Autoencoder

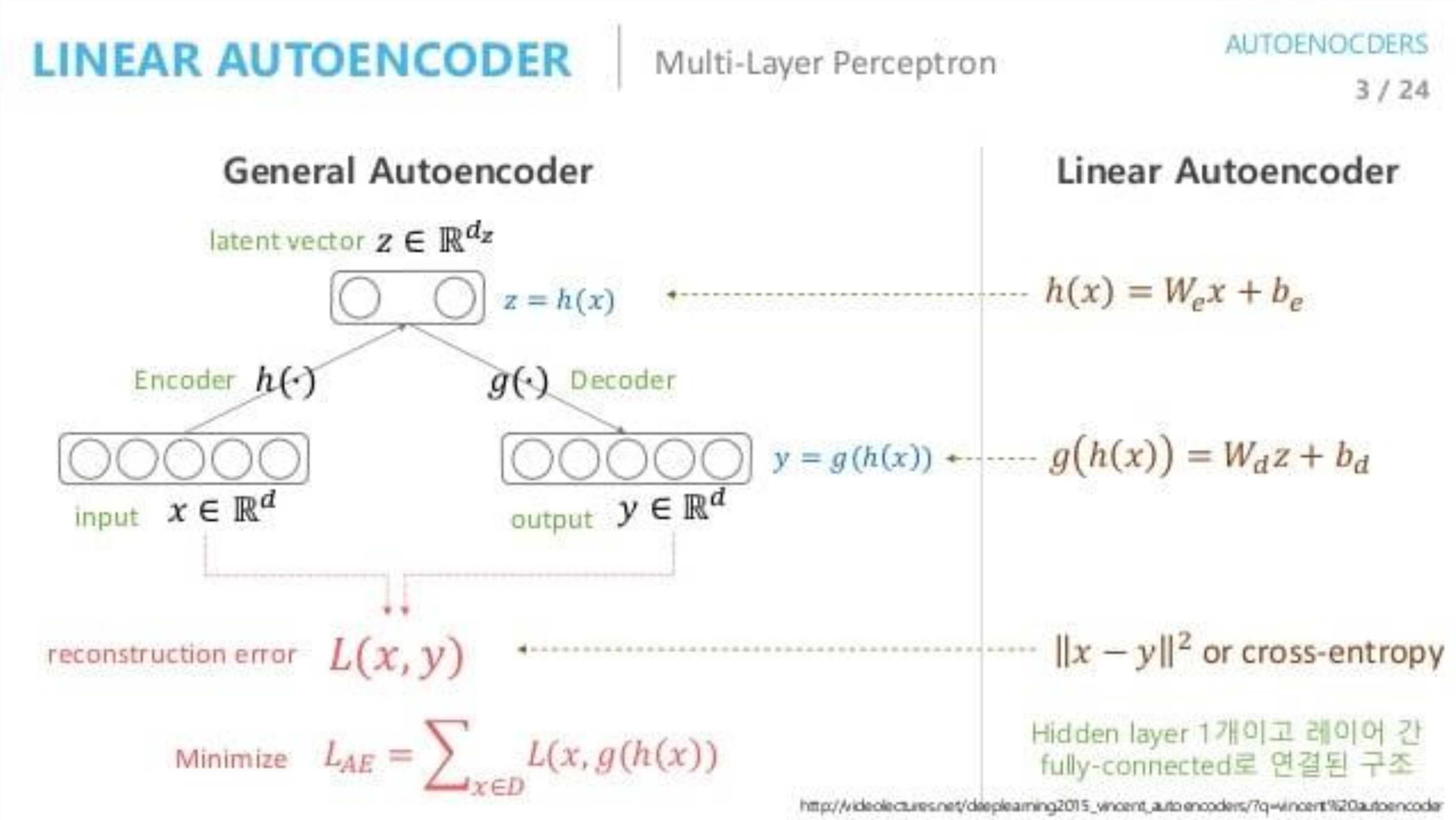
Trained to reconstruct input x as output $\tilde{x}$

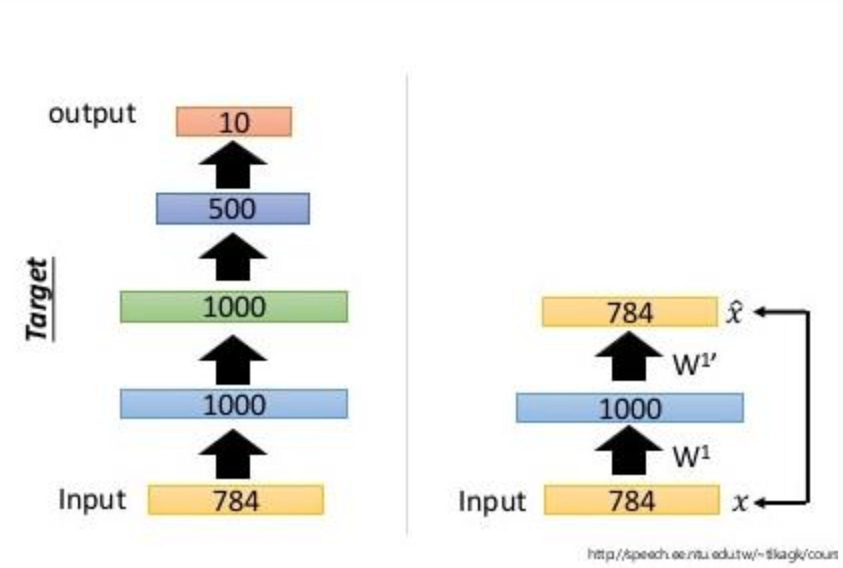


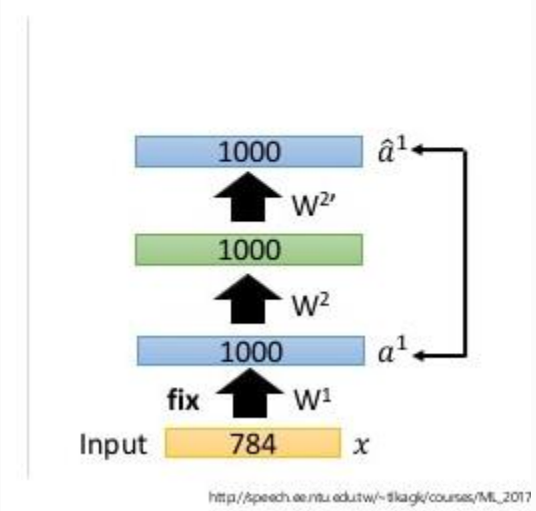
Unsupervised Learning → Supervised Learning (Self Learning)

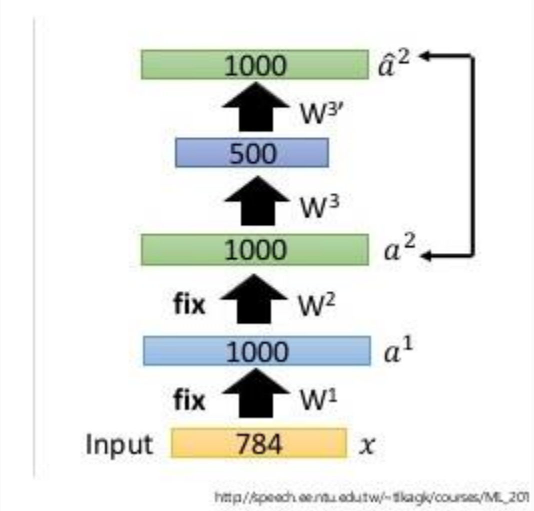
(+ minimum 성능 보장)

Linear Autoencoder = PCA





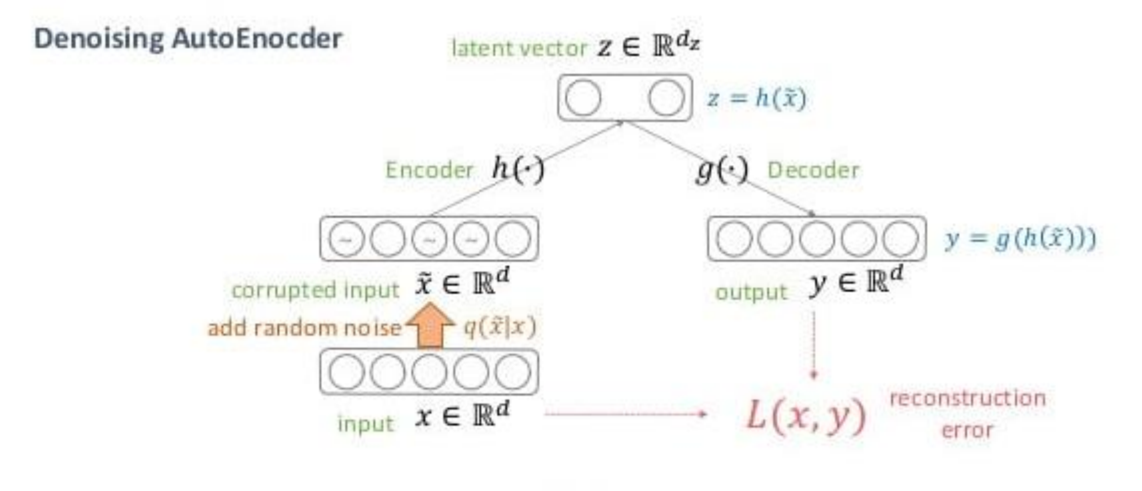




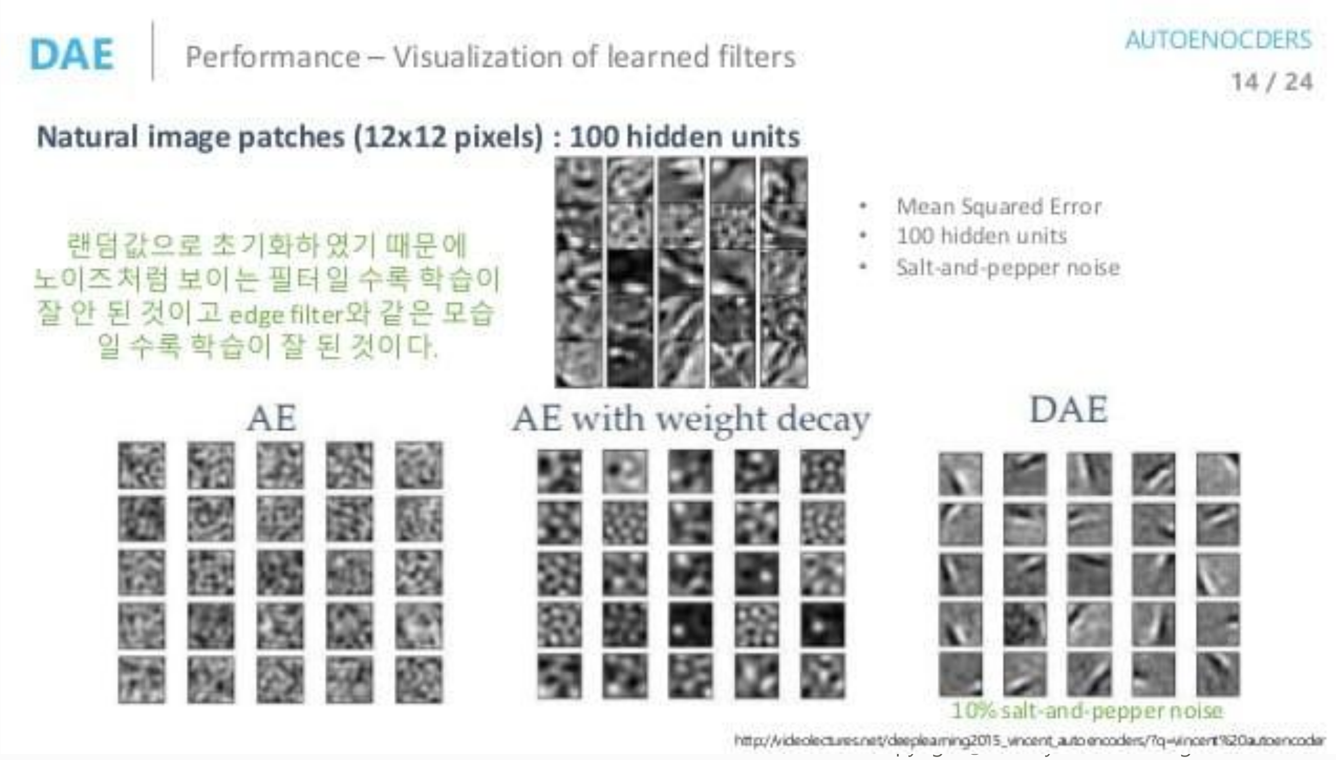
Tied weighted.

W' W = I

Denoising Autoencoder

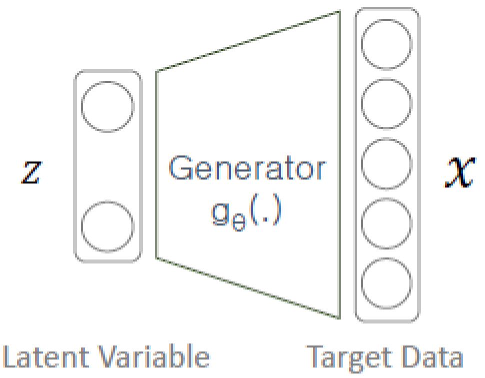


Noise가 끼어 있지만 의미적으로 같아야 한다
(manifold 위에서 같은 점에 위치)



Variational Autoencoders

KEY WORD : Generative model
(Autoencoder : manifold learning)



$z \sim p(z)$	Random variable
$g_\theta(\cdot)$	Deterministic function parameterized by θ
$x = g_\theta(z)$	Random variable

Latent variable can be seen as a set of control parameters for target data (generated data)

For MNIST example, our model can be trained to generate image which match a digit value z randomly sampled from the set [0, ..., 9].

그래서, p(z)는 샘플링 하기 용이해야 편하다.

$$p(x|g_\theta(z)\ ) = p_\theta(x|z)$$

We are aiming maximize the probability of each x in the training set, under the entire generative process, according to:

$$\int p(x|g_\theta(z))p(z)dz = p(x)$$

Normal Distribution?

Question: Is it enough to model p(z) with simple distribution like normal distribution?

Yes

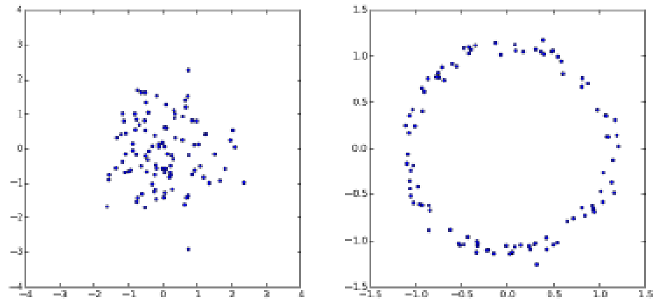


Figure 2: Given a random variable z with one distribution, we can create another random variable $X = g(z)$ with a completely different distribution. Left: samples from a gaussian distribution. Right: those same samples mapped through the function $g(z) = z/10 + z/||z||$ to form a ring. This is the strategy that VAEs use to create arbitrary distributions: the deterministic function g is learned from data.

Tutorial on Variational Autoencoders : https://arxiv.org/pdf/1606.05908

VAE

Question: Why don't we use maximum likelihood estimation directly?

$$p(x) \approx \sum_i p(x|g_\theta(z_i))p(z_i)$$

If $p(x|g_\theta(z)) = \mathcal{N}(x|g_\theta(z), \sigma^2 * I)$, the negative log probability of X is proportional squared Euclidean distance between $g_\theta(z)$ and x.

x : Figure 3(a)
$z_{bad} \rightarrow g_\theta(z_{bad})$: Figure 3(b)
$z_{bad} \rightarrow g_\theta(z_{good})$: Figure 3(c) (identical to x but shifted down and to the right by half a pixel)

$$\|x - z_{bad}\|^2 < \|x - z_{good}\|^2 \rightarrow p(x|g_\theta(z_{bad})) > p(x|g_\theta(z_{good}))$$

Solution 1: we should set the σ hyperparameter of our Gaussian distribution such that this kind of erroroneous digit does not contribute to p(X) → hard..
Solution 2: we would likely need to sample many thousands of digits from z_{good} → hard..

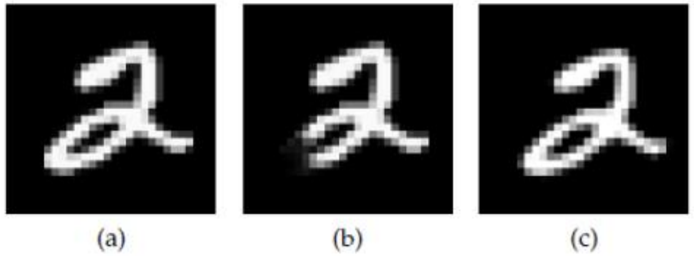


Figure 3: It's hard to measure the likelihood of images under a model using only sampling. Given an image X (a), the middle sample (b) is much closer in Euclidean distance than the one on the right (c). Because pixel distance is so different from perceptual distance, a sample needs to be extremely close in pixel distance to a datapoint X before it can be considered evidence that X is likely under the model.

생성기에 대한 확률모델을 가우시안으로 할 경우, MSE관점에서 가까운 것이 더 p(x)에 기여하는 바가 크다.
MSE가 더 작은 이미지가 의미적으로도 더 가까운 경우가 아닌 이미지들이 많기 때문에 현실적으로 올바른 확률값을 구하기가 어렵다.

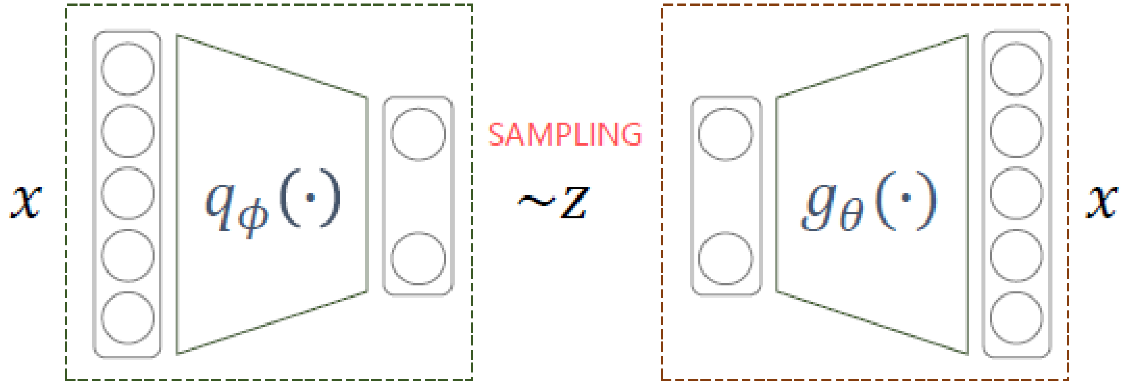
Tutorial on Variational Autoencoders : https://arxiv.org/pdf/1606.05908

Loss Function of VAE

Relationship among $p(x), p(z|x), q_\phi(z|x)$

LOSS FUNCTION | NeuralNet Perspective

$$\arg\min_{\phi,\theta} \sum_i \underbrace{-\mathbb{E}_{q_\phi(z|x_i)}\left[\log\left(p(x_i|g_\theta(z))\right)\right] + KL\left(q_\phi(z|x_i)||p(z)\right)}_{L_i(\phi,\theta,x_i)}$$



$q_\phi(z|x)$ | $g_\theta(x|z)$

Encoder
Posterior
Inference Network

Decoder
Generator
Generation Network

The mathematical basis of VAEs actually has relatively little to do with classical autoencoders

Tutorial on Variational Autoencoders : https://arxiv.org/pdf/1606.05908

Loss Function of VAE

LOSS FUNCTION | Explanation

$$\underset{\phi,\theta}{\arg\min} \sum_i \underbrace{-\mathbb{E}_{q_\phi(z|x_i)}\left[\log\left(p(x_i|g_\theta(z))\right)\right] + KL\left(q_\phi(z|x_i)||p(z)\right)}_{L_i(\phi,\theta,x_i)}$$

원 데이터에 대한 likelihood

Variational inference를 위한 approximation class 중 선택

다루기 쉬운 확률 분포 중 선택

$$L_i(\phi,\theta,x_i) = \underbrace{-\mathbb{E}_{q_\phi(z|x_i)}\left[\log\left(p(x_i|g_\theta(z))\right)\right]}_{\text{Reconstruction Error}} + \underbrace{KL\left(q_\phi(z|x_i)||p(z)\right)}_{\text{Regularization}}$$

Reconstruction Error

- 현재 샘플링 용 함수에 대한 negative log likelihood
- x_i에 대한 복원 오차 (AutoEncoder 관점)

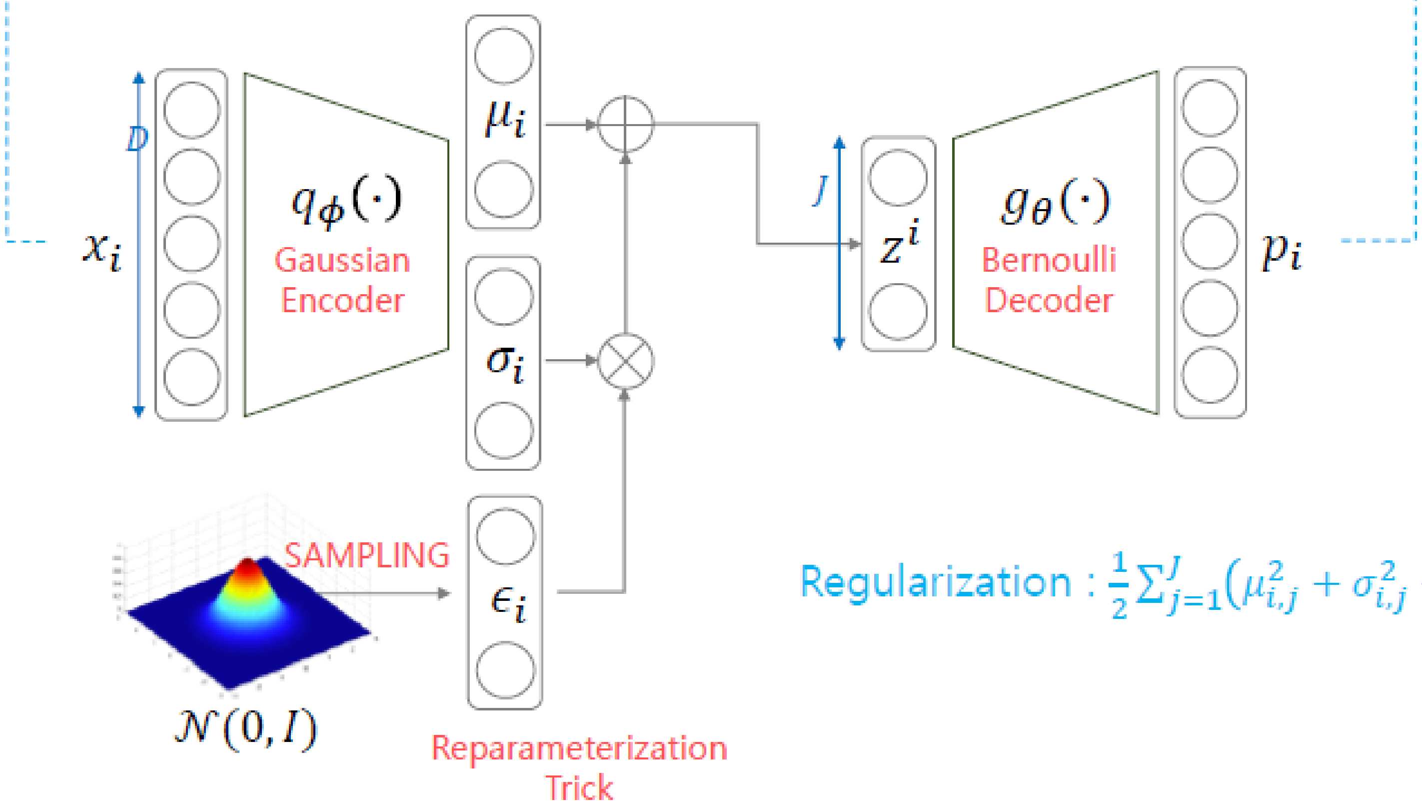
Regularization

- 현재 샘플링 용 함수에 대한 추가 조건
- 샘플링의 용의성/생성 데이터에 대한 통제성을 위한 조건을 prior에 부여 하고 이와 유사해야 한다는 조건을 부여

Loss Function of VAE

STRUCTURE | Default : Gaussian Encoder + Bernoulli Decoder

Reconstruction Error: $-\sum_{j=1}^{D} x_{i,j} \log p_{i,j} + (1 - x_{i,j}) \log(1 - p_{i,j})$



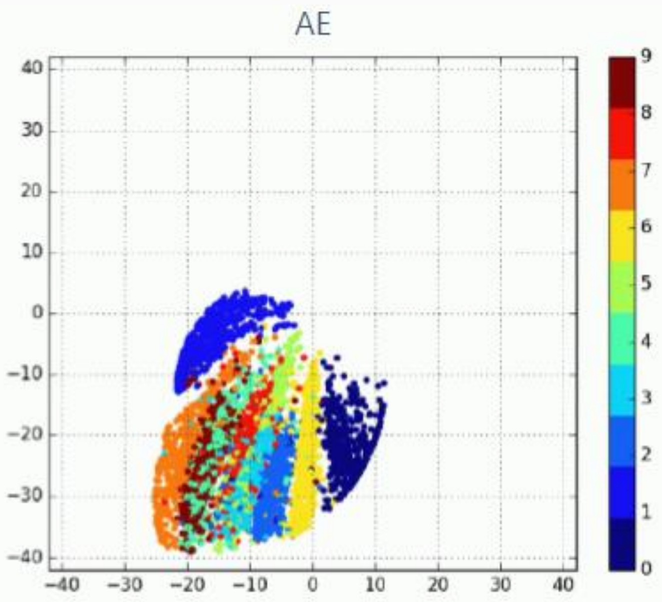
Regularization : $\frac{1}{2}\sum_{j=1}^{J}(\mu_{i,j}^2 + \sigma_{i,j}^2 - \ln(\sigma_{i,j}^2) - 1)$

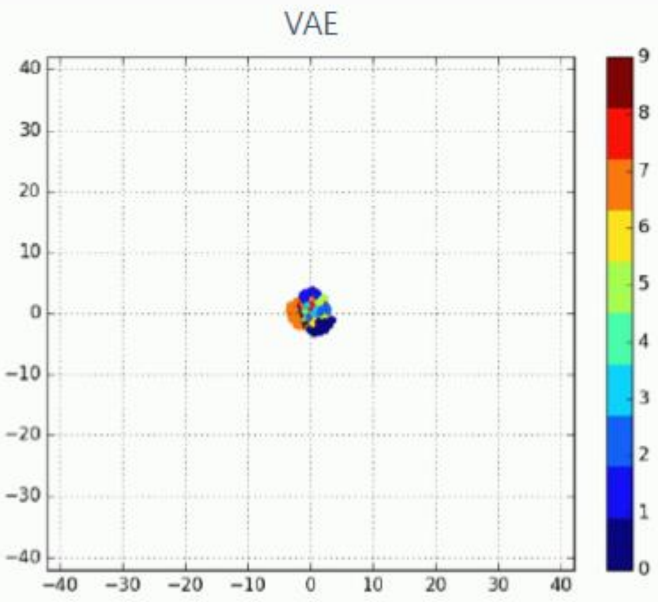
AE vs VAE

AE & VAE 코드 관점에서 한 줄 다름.
(Loss function 에서 KL_divergence term 추가)

AE – for dimensionally reduction. Generating에는 적합하지 않음.
VAE – for generating. Manifold 위치가 안정적임.

Reconstruct만 할거면 AE Loss가 훨씬 저렴함.
AE는 text, image, sound, …, multi_model까지 domain을 가리지 않고 안정적이라고 함.





영상 자료 : https://www.youtube.com/watch?v=rNh2CrTFpm4
43분 15초~

Adversarial Autoencoder (AAE)

Adversarial Autoencoder

$$L_i(\phi,\theta,x_i) = -\mathbb{E}_{q_\phi(z|x_i)}\left[\log\left(p_\theta(x_i|z)\right)\right] + \boxed{KL(q_\phi(z|x_i) \parallel p(z))}$$

Regularization

Conditions for $q_\phi(z|x_i)$, $p(z)$

1. Easily draw samples from distribution
2. KL divergence can be calculated

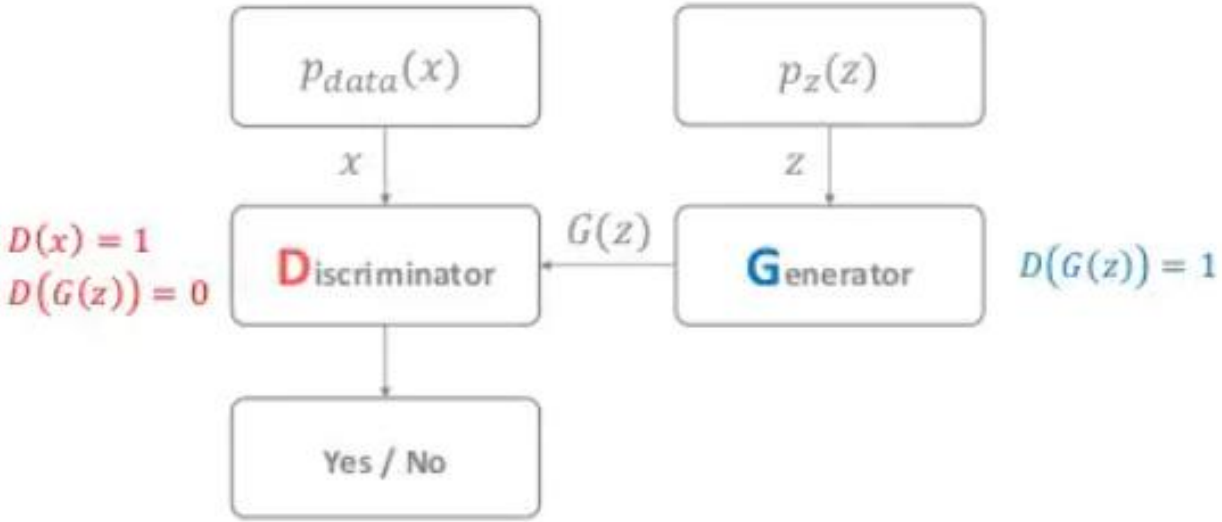
Adversarial Autoencoder (AAE)

Conditions	$q_\phi(z\|x_i)$	$p(z)$
Easily draw samples from distribution	O	O
KL divergence can be calculated	X	X

KL divergence is replaced by discriminator in GAN

AAE

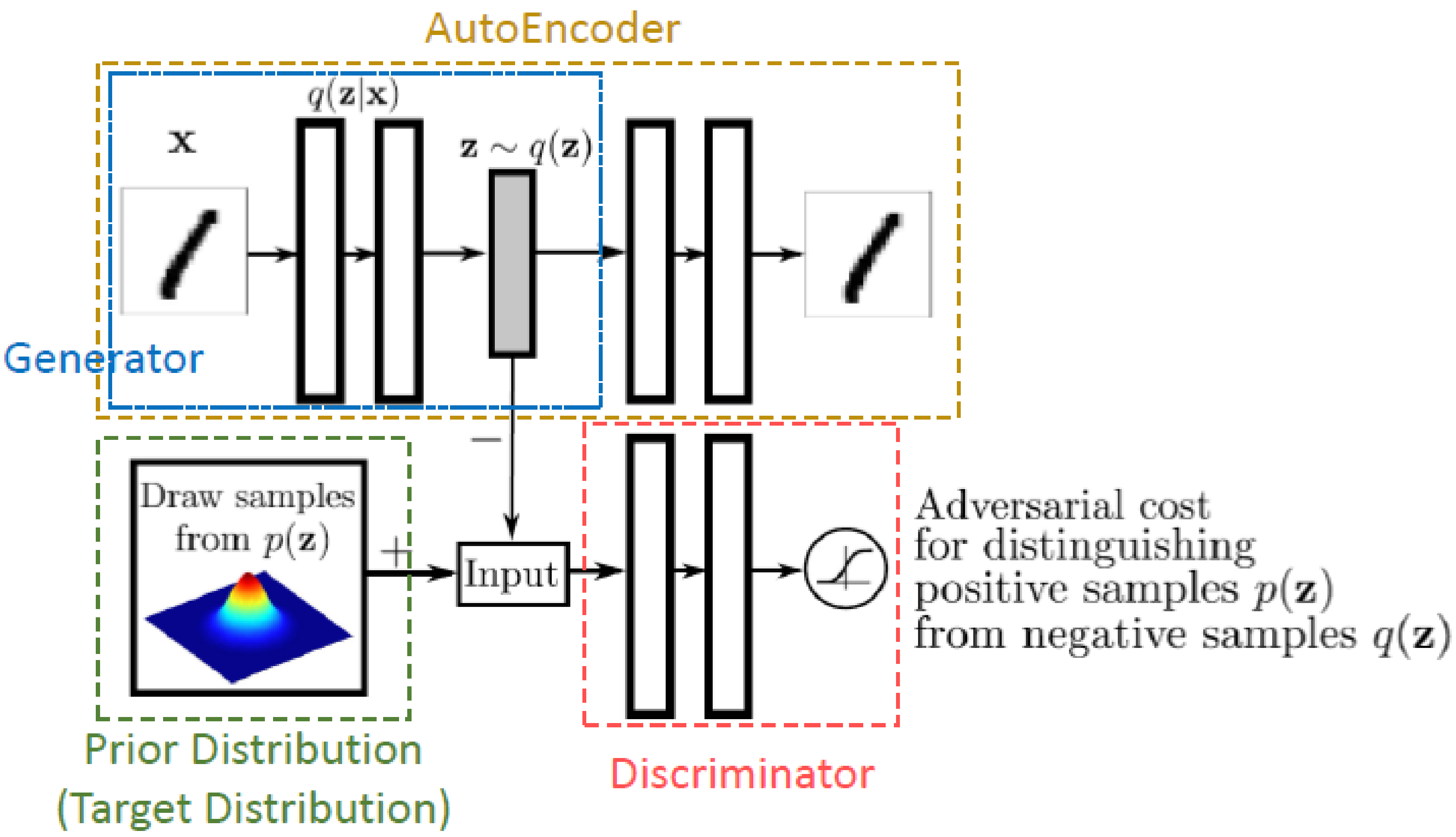
Generative Adversarial Network



Value function of GAN : $V(D,G) = \mathbb{E}_{x \sim p_{data}(x)}[\log D(x)] + \mathbb{E}_{z \sim p_z(z)}\left[\log\left(1 - D(G(z))\right)\right]$

Goal : $D^*, G^* = \min_G \max_D V(D,G)$ GAN은 $G(z) \sim p_{data}(x)$로 만드는 것이 목적이다

AAE


AutoEncoder
$q(\mathbf{z}|\mathbf{x})$
$\mathbf{x}$
$\mathbf{z} \sim q(\mathbf{z})$
Generator
−
Draw samples from $p(\mathbf{z})$
+
Input
Adversarial cost for distinguishing positive samples $p(\mathbf{z})$ from negative samples $q(\mathbf{z})$
Prior Distribution (Target Distribution)
Discriminator

AAE

Loss Function

GAN loss

$$V(D,G) = \mathbb{E}_{z\sim p(z)}[\log D(z)] + \mathbb{E}_{x\sim p(x)}\left[\log\left(1 - D\left(q_\phi(x)\right)\right)\right]$$

Let's say G is defined by $q_\phi(\cdot)$ and D is defined by $d_\lambda(\cdot)$

$$V_i(\phi,\lambda,x_i,z_i) = \log d_\lambda(z_i) + \log\left(1 - d_\lambda\left(q_\phi(x_i)\right)\right)$$

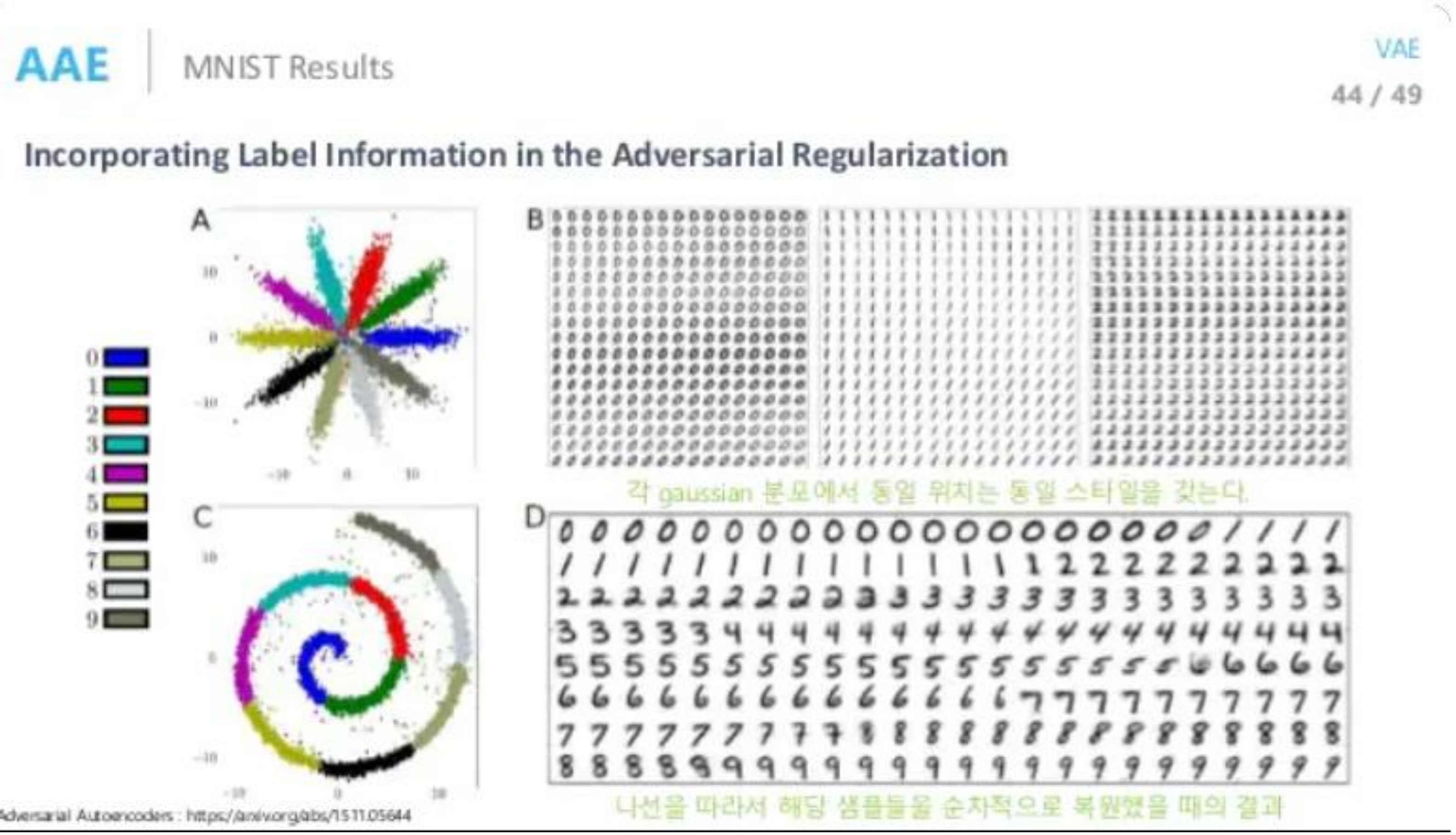
*논문에는 로스 정의가 제시되어 있지 않아 새로 정리한 내용

VAE loss

$$L_i(\phi,\theta,x_i) = -\mathbb{E}_{q_\phi(z|x_i)}[\log(p_\theta(x_i|z))] + \cancel{KL(q_\phi(z|x_i)||p(z))}$$

AAE

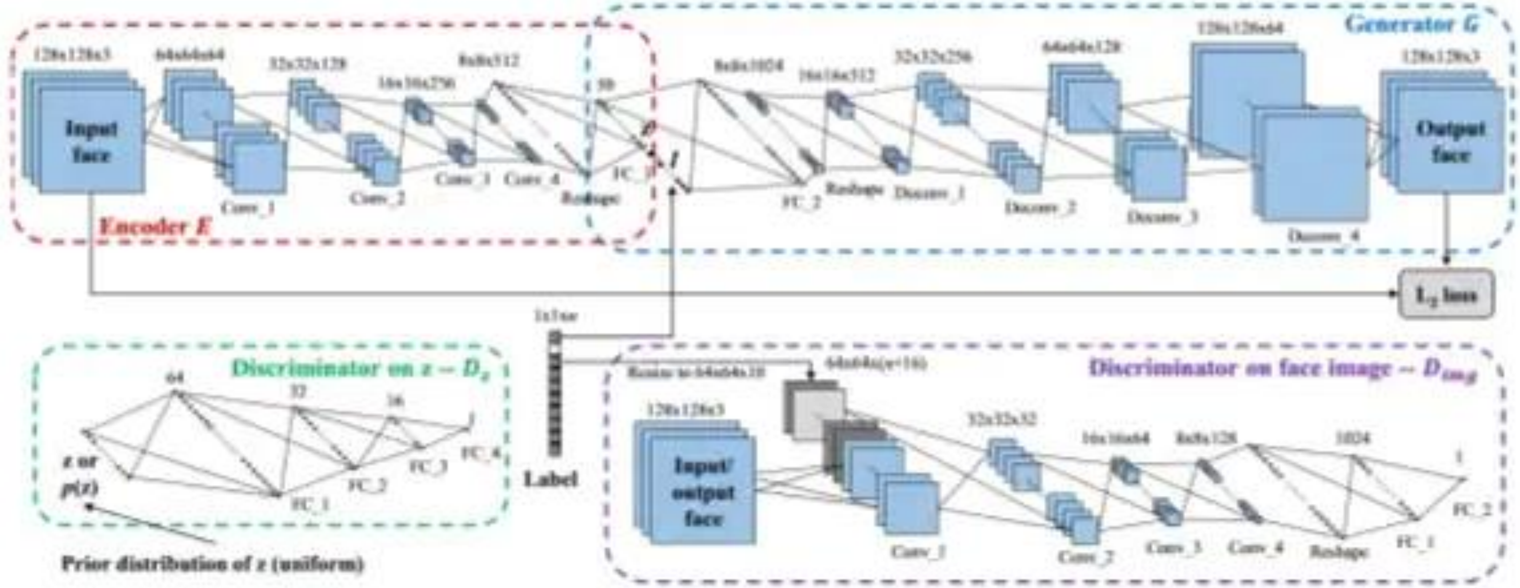
AAE는 prior distribution을 원하는 모양으로 만들 수 있음.



GAN + VAE…?

GAN+VAE | Papers in CVPR2017

Age Progression/Regression by Conditional Adversarial Autoencoder



https://zzutk.github.io/Face-Aging-CAAE/

Q & A

Reference

-오토인코더의 모든 것(이활석 NAVER)
(https://www.slideshare.net/NaverEngineering/ss-96581209)
-수학의 즐거움 정보기하
(https://www.youtube.com/watch?v=4s06EgHHRrA)