Auto-Encoding Variational Bayse

Kingma and Welling. "*Auto-Encoding Variational Bayes*, *International Conference on Learning Representations." ICLR,* 2014.
arXiv:1312.6114

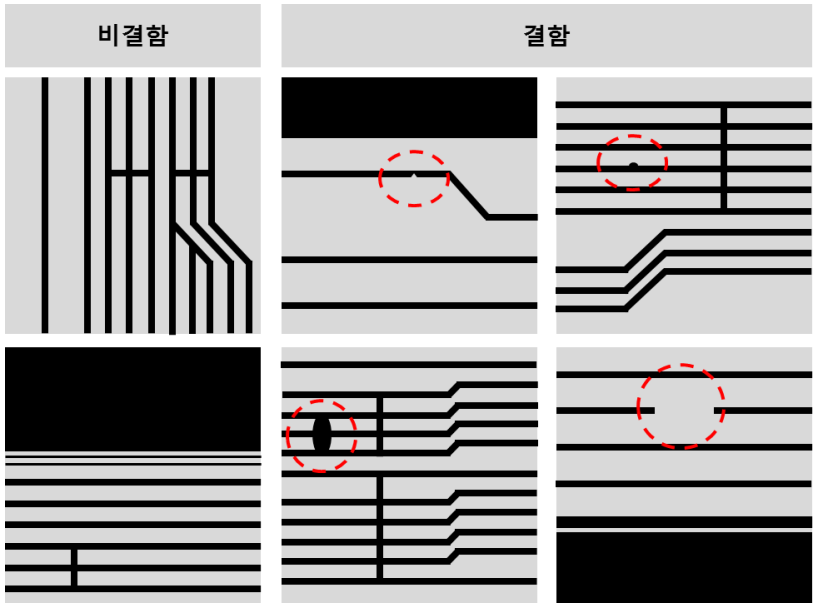
목차

- Generative Model
- Variational Inference
- Variational AutoEncoder (VAE)

Motivation

- 현업데이터(검사 도메인) controllable 한 EDA를 하기 위함
- 현업데이터 가능하다면 Generation
- 여러 논문들에서 중요한 개념이 되는 것을 공부하고자

현업데이터? Ambiguous? Model Capacity?



What is Deep Generative Model

Learning a Probability Distribution P(X)

- Generative Model의 궁극적목적:
  - 데이터가생성되는과정,즉확률분포P(X) 를학습하고싶은것.

- 어떤변수X의확률분포P(X)를안다는것은그변수에대한 모든것을안다는의미다.

- 확률분포P(X)를알고안다면,크게2가지를할수있다.

  1. (Inference) X의내부적 생성구조 , latent code z혹은class Y정보에 관한추론
  2. (Sampling) Data 생성 X ~ P(X)

- Generative Model을학습하는 유일한 이유.

확률모형을 어떻게 학습 하지?

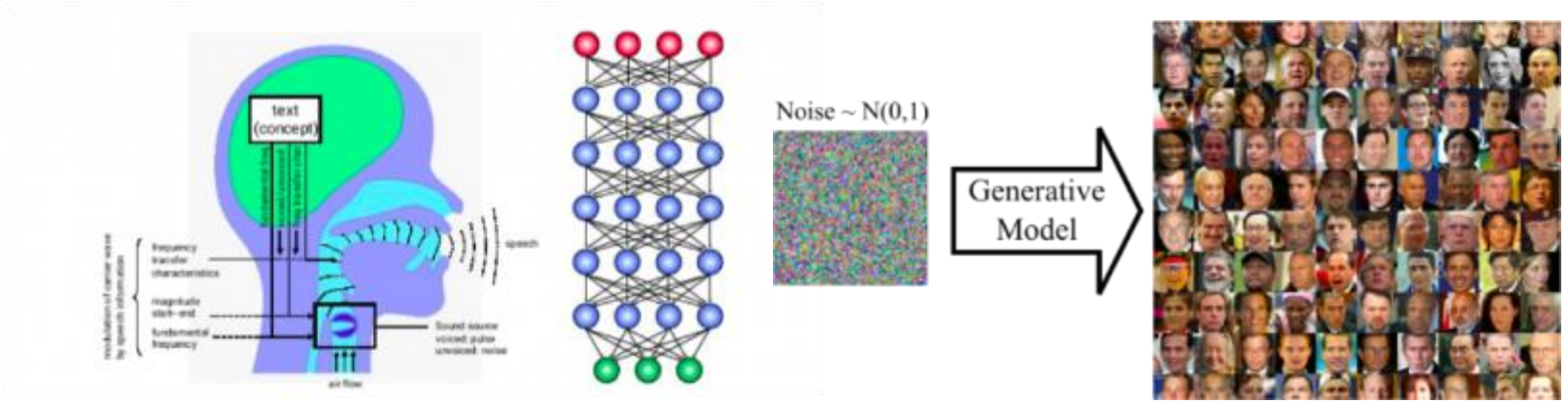
- 다양한 **Generative Model**
  - Naïve Bayes, GDA, GMM, Beta-Bernoulli 모형…
  - 이밖에 LDA (Topic Model), Hidden Markov Mode, non-parametric LDA, Markov Random Fields등 다양한 전통적 Generative Model들이있다.

- **Generative Model을 학습(parameter를추정)하는 몇가지 방법론 기법**

  - MLE, MAP, EM, Analytical Posterior Inference, **(Stochastic) variational Inference w/ Re-parameterization Trick..**

- 대개 주어진데이터에 대한 확률분포P(X), 혹은 숨겨진구조 latent variable Z에대한Prior P(Z)등을 가정하고, **주어진데이터를 잘 설명 하도록 모형을 학습했다.**

전통적인Generative Model의한계



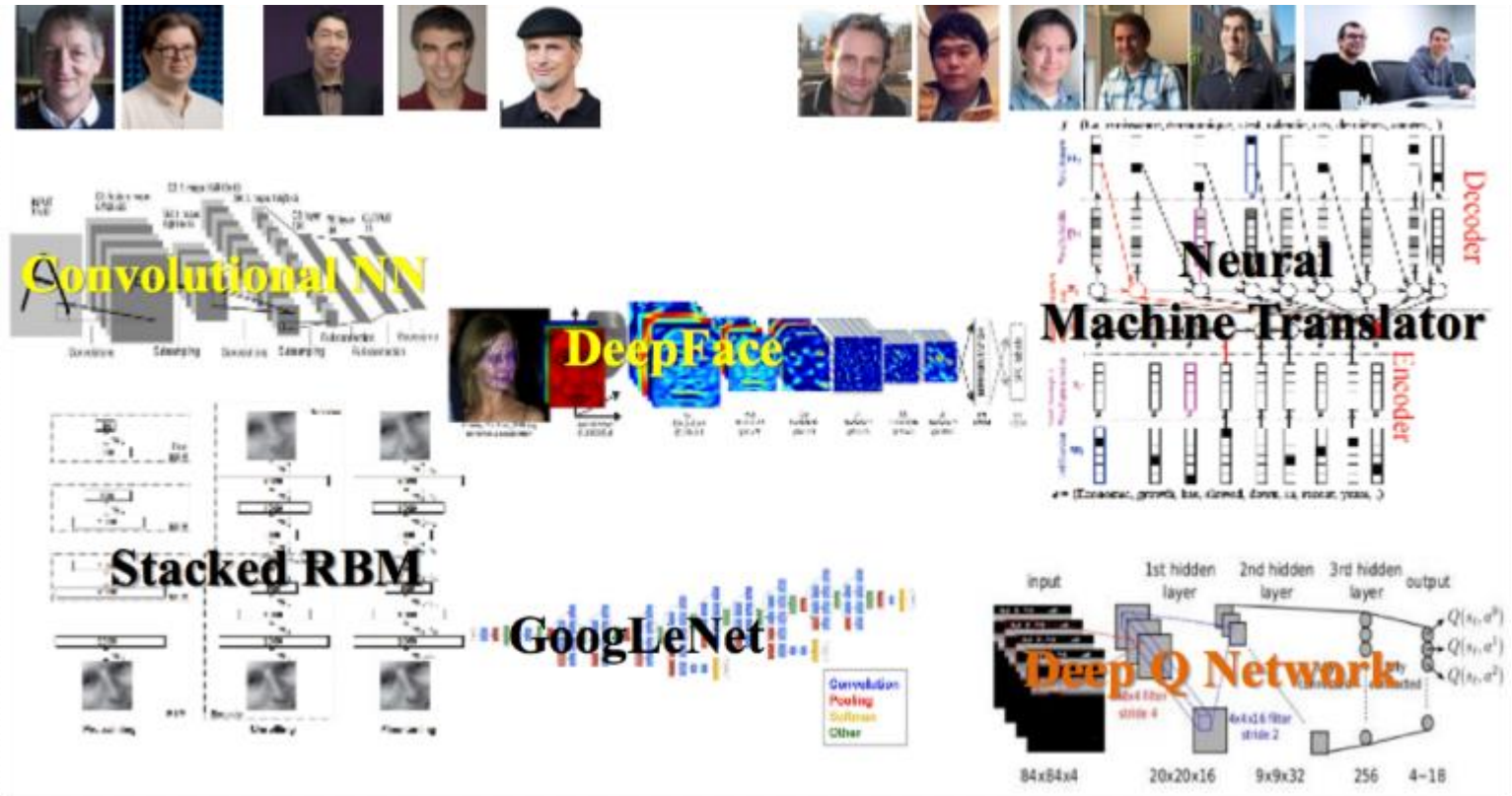
- 기존인풋데이터에대한식별및추론에초점을두었던것과는달리, 딥러닝의시대에서는데이터의생성이중요한task다.

- 우리는인공지능이이미지,언어,소리등을만들어내길원한다.

- 하지만,현실세계의고차원적 데이터생성task를실현하기에는기존 확률모형들의표현력및학습방법등에분명한한계가있다.

Deep Learning & Supervised Learning

- 한편,Supervised Learninig 영역에서는인공신경망과 같이 아주복잡도가높은모형들에대한연구가많이이루어졌다.

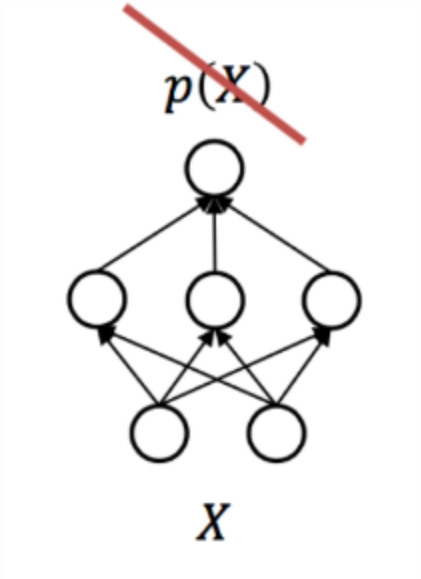


What is Deep Generative Model

"그렇다면..NN을 이용해서 확률모
형을학습할수는없을까?"

나이브한 접근

- 확률X의 확률밀도를 학습한다?



- 각 학습데이터에 대한 확률밀도값이 주어진다면 **지도학습 (Supervised Learning)**을시도할수있다. 그러나,일반적으로 데이터 가많아도 각 데이터의 정확한 확률밀도값을 알수없기 때문에감독 학습만으로는 정확한 확률밀도값을 추정하기 힘들다.

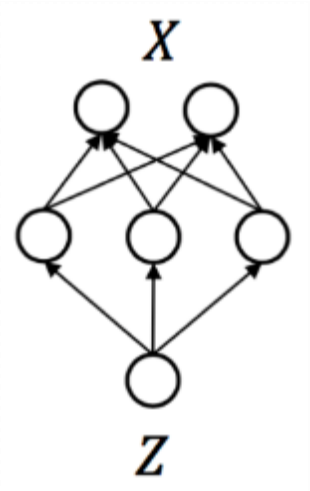
Deep Neural Network를데이터Sampling?

- 확률값추론보다는데이터를잘샘플링(생성)하는모형을만들어 보면어떨까?
- 인공신경망은한벡터를 다른벡터로 변환하는것 을 아주잘한다.

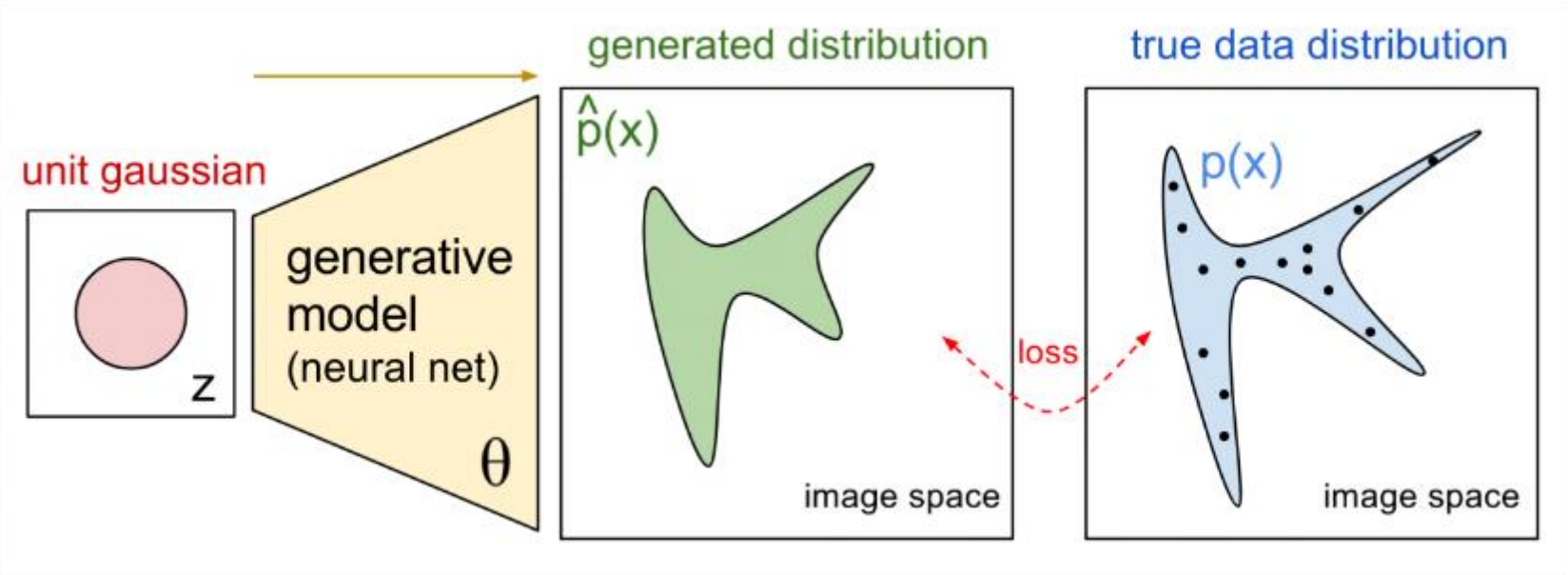
Key아이디어는

1. 간단한 확률 분포에서 얻어진 벡터를 z~P(z) e.g. Gaussian
2. Neural Network를통해서실제와비슷한데이터X로변환해보자!

$$Z \sim N(0, \mathbf{I})$$
$$X = f(Z)$$

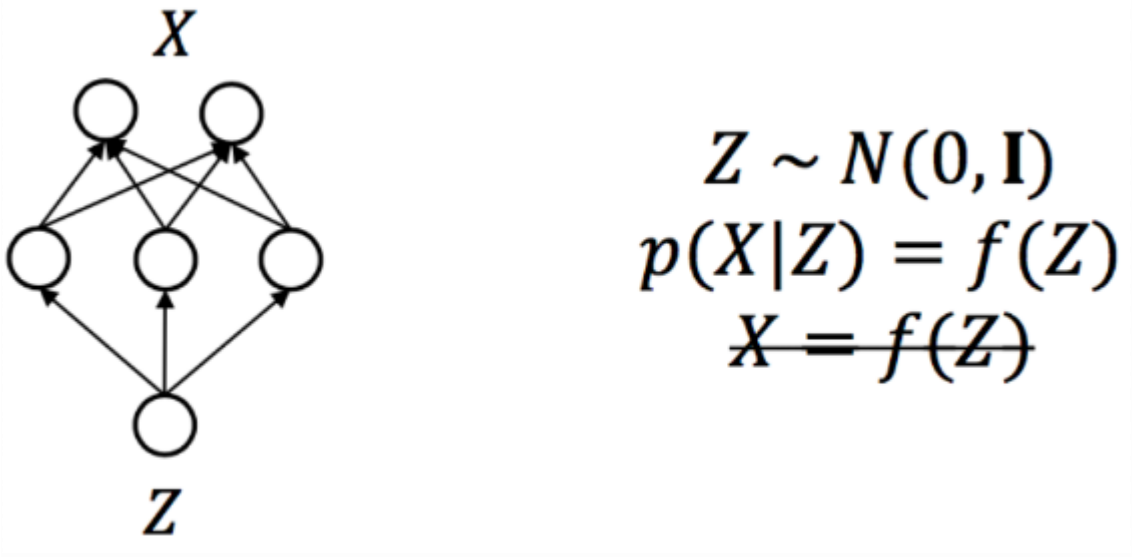


Deep Generative Model



Probability Distribution meets Deep Learning

- Random Variable의 함수를 적용하여 얻은 변수
  ==새로운 Random Variable

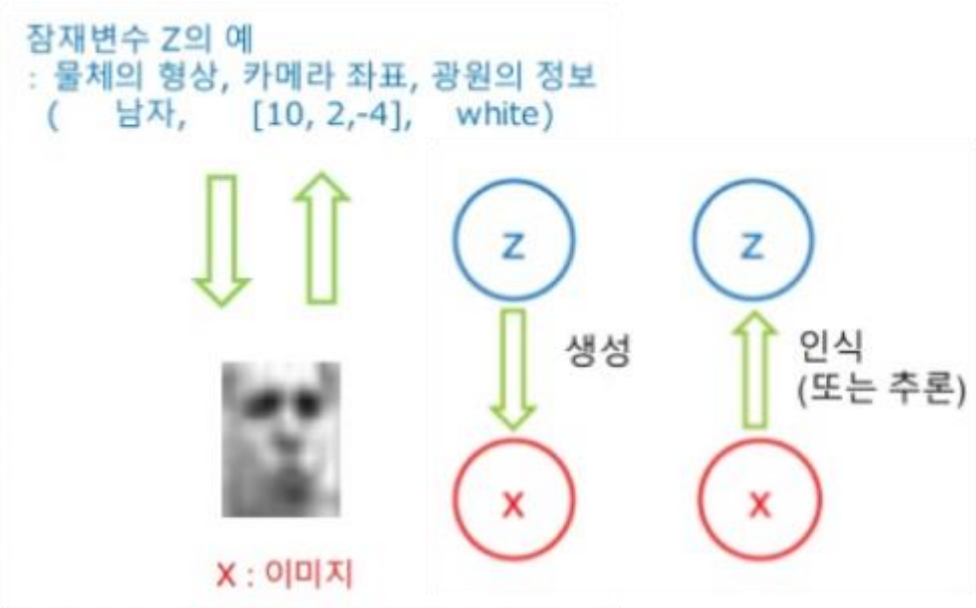


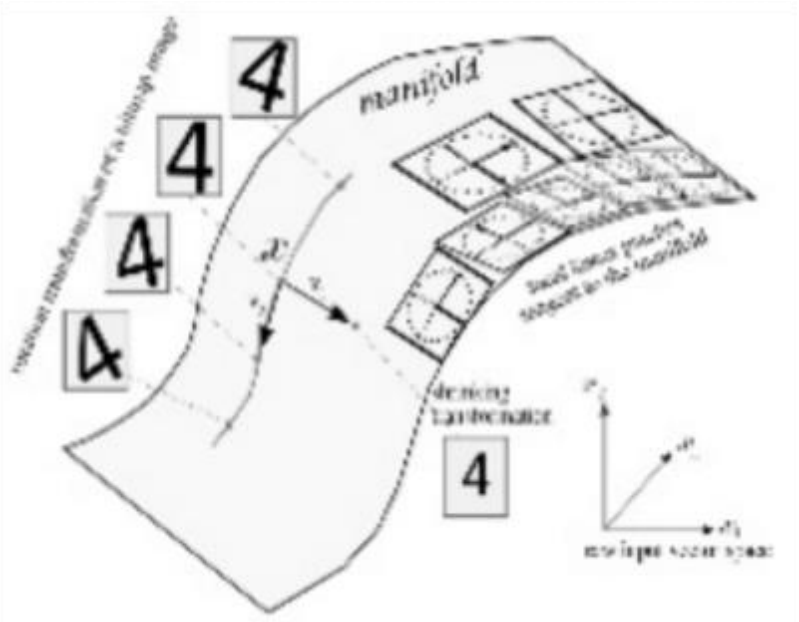
$$Z \sim N(0, \mathbf{I})$$
$$p(X|Z) = f(Z)$$
$$\sout{X = f(Z)}$$

- 엄밀하게말하자면인공신경망을이용해서Conditional Probability P(X|Z)를모델링하자는것이다.

Latent Space? Z? Manifold?

- 우리가관측한데이터x는고차원이지만,실제의미가표현되고데이터의생성에관여하는latent variable z 는저차원으로표현되고, 이 저차원상에 비슷한데이터가 군집하는 공간이있을것이다.

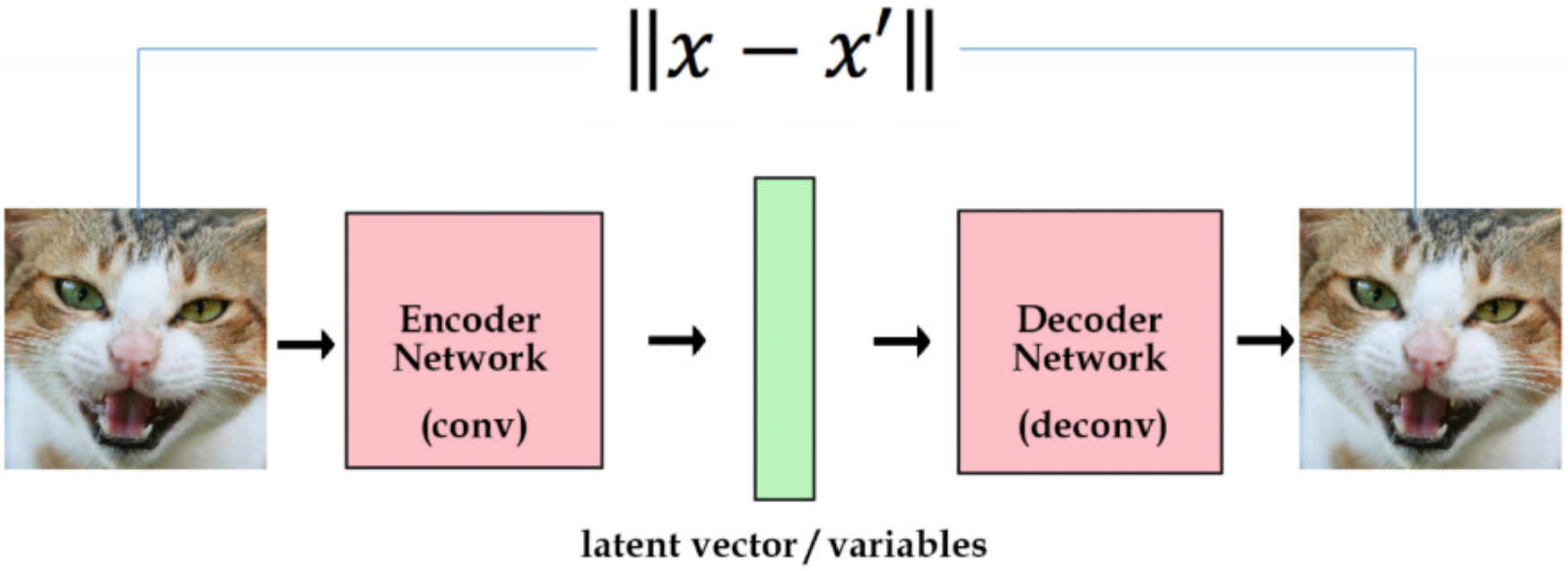




Manifold hypothesis

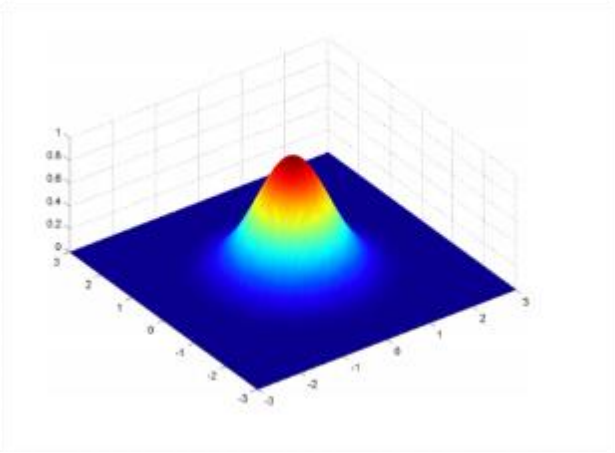
(vanilla) Auto-Encoder

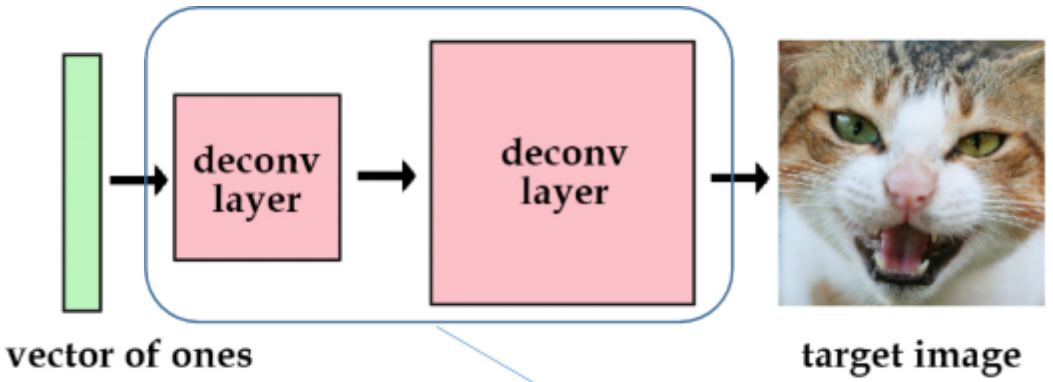
- NN를 이용해서 각 영상에 대한 Latent Vector z를나타내는방법 및 z를사용해서 영상을 복원하는방법을 동시에 학습.
- 원본데이터x 자체가일종의 레이블역할을하여reconstructed된 데이터또는decoded 데이터와의 차이로 loss를계산.
- L1 loss나L2 loss를많이사용함(MSE)
- Unsupervised Learning, Feature Learning, Dimensionality Reduction



Variational Auto-Encoder의 목표

- (영상)데이터를잘생성하는확률모형P(x|z;)을학습하는것.
- z~P(Z)는Gaussian 분포사용.

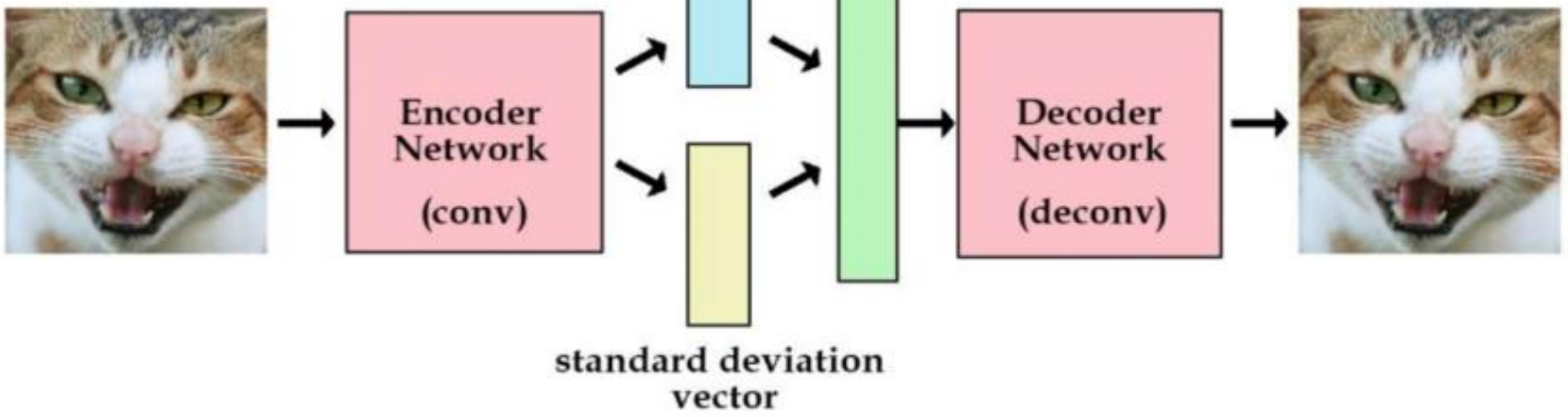




$$P(z) = \mathcal{N}(z|0, I)$$

$$P(X|z;\theta) = \mathcal{N}(X|f(z;\theta), \sigma^2 * I)$$

$q(z|x)$ $p(x|z)$



$$l_i(\theta, \phi) = -E_{z \sim q_\theta(z|x_i)}[\log p_\phi(x_i|z)] + KL(q_\theta(z|x_i)||p(z))$$

Reconstruction Loss

$\|x - x'\|$

KL Divergence
Regularizer

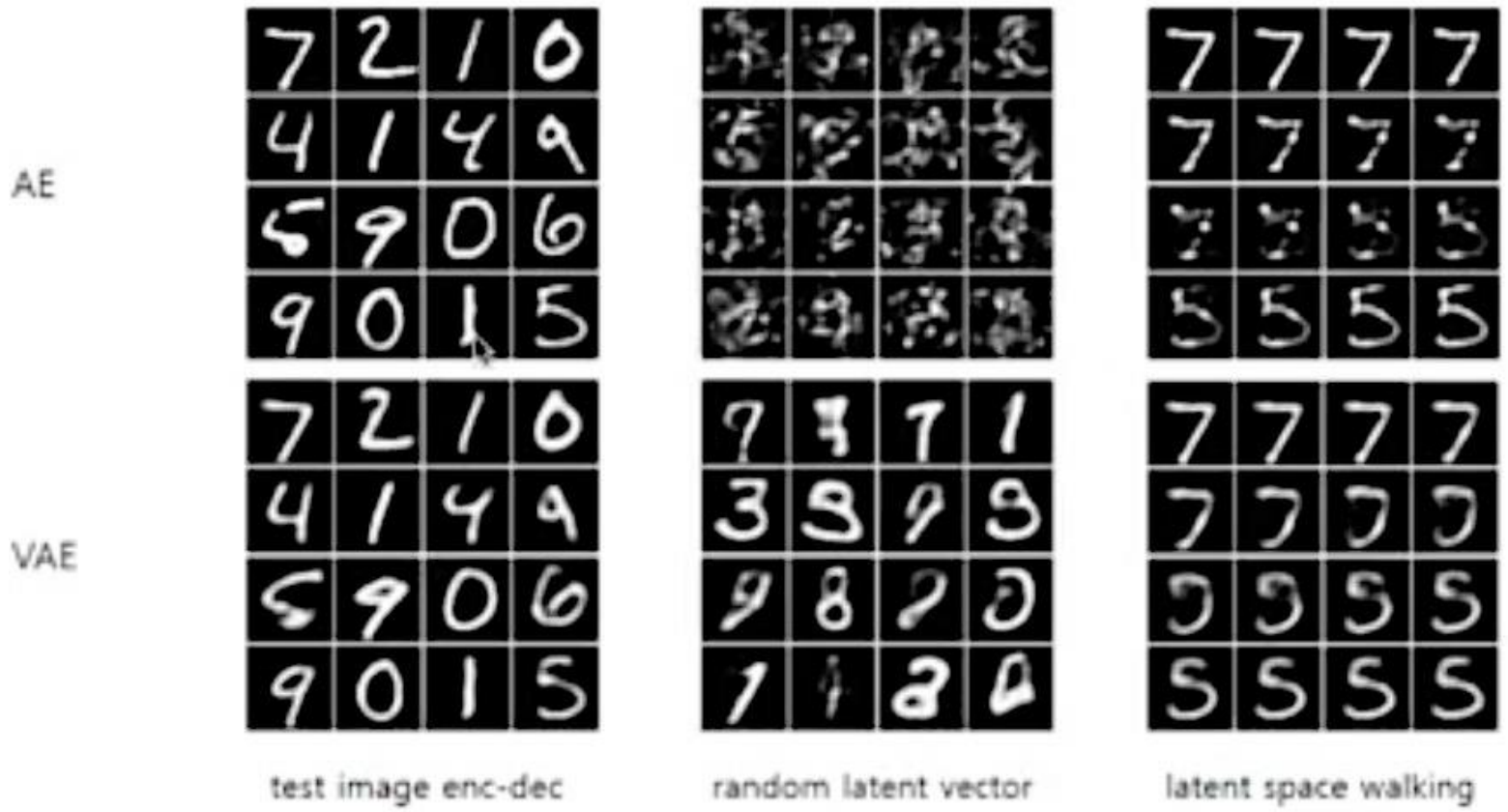
AE vs VAE

- VAE는AE와다르게 latent vector z에대한 KL divergence **Regularization Term**이 추가적으로 붙는다.

- VAE는 AE와다르게 **z를중간에 Sampling**하여 데이터x를 복원하는방식으로 학습한다. 이러한방식은 Test-time에 새로운데이터x를 생성(샘플링)할수있게한다(= x에대한확률분포를학습한것).

- 데이터x를 latent space z의차원으로Mapping하는방법, Encoder q(z|x)에 대한 확률모형을 명시적으로얻을수 있다(= Posterior Distribution).

AE vs VAE차이비교



Backgroud Keywords

Variational Inference (Bayesian Inference)

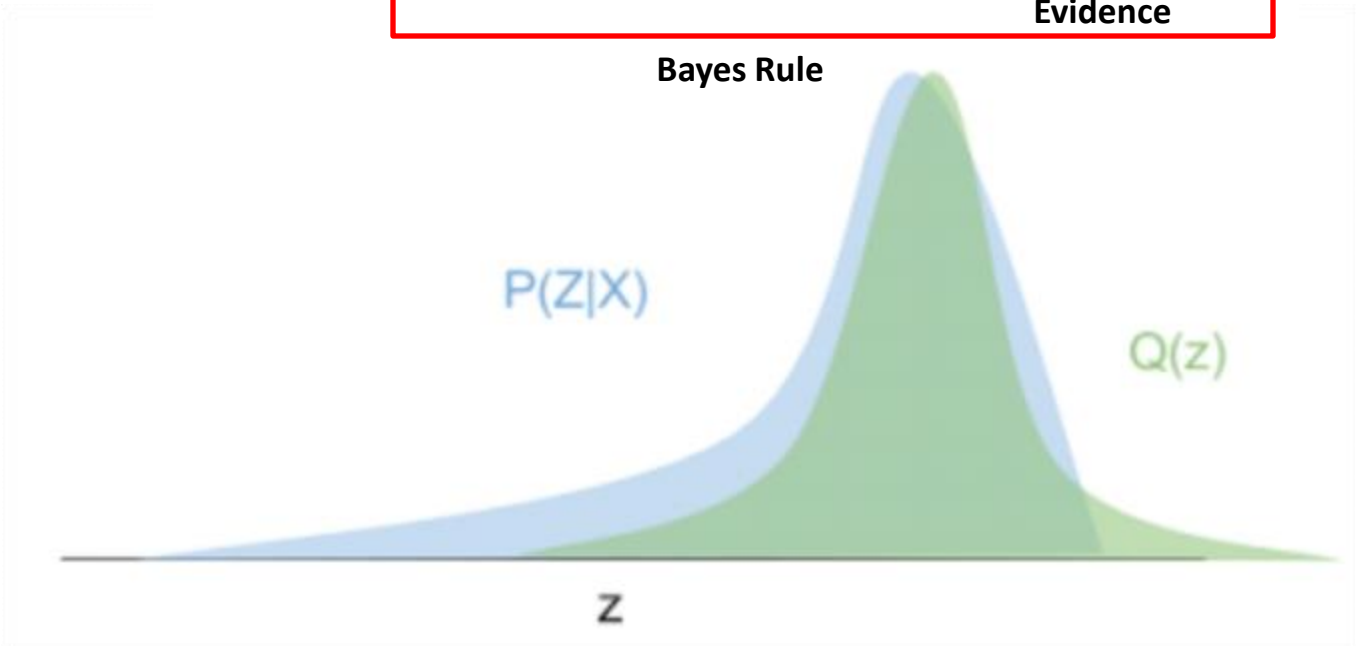
변분추론(VariationalInference)의 목적

- Posterior를 다루기쉬운분포 Q(z)로 근사 하는것임 (KLD 이용)

$$Q(z) \approx P(z|x) = \frac{P(x|z)P(z)}{P(x)} = \frac{P(x|z)P(z)}{\sum_z P(x,z)}$$

Evidence

Bayes Rule



VI를 껴넣으면 모델 어떻게 학습?
=> KL Divergence 관점

- KL divergence를 이용해서 q(z)와 p(z|x)와 오차를 측정!

$$D_{\mathrm{KL}}\left(q(z)||p(z\mid \boldsymbol{x})\right) = \int q(z)\log\frac{q(z)}{p(z\mid \boldsymbol{x})}dz$$

gaussian

$$= \int q(z)\log\frac{q(z)p(x)}{p(\boldsymbol{x}\mid z)p(z)}dz$$

$$= \int q(z)\log\frac{q(z)}{p(z)}dz + \int q(z)\log p(\boldsymbol{x})dz - \int q(z)\log p(\boldsymbol{x}\mid z)dz.$$

$$= D_{\mathrm{KL}}\left(q(z)||p(z)\right) + \log p(\boldsymbol{x}) - \mathbb{E}_{z\sim q(z)}\left[\log p(\boldsymbol{x}\mid z)\right].$$

- KL divergence term이 줄어들때 까지 $q(z)$ 를 학습하면 됨

- **결론:** $p(z|\boldsymbol{x})$를 잘 근사한 $q^*(z)$를 얻음
  - 원래 알고싶었던 이놈(Posterior)
  - Latent effective knowledge set.
  - posterior

Evidence Lower Bound (ELBO) 관점

$$D_{\mathrm{KL}}\left(q(z)||p(z\mid \boldsymbol{x})\right) = D_{\mathrm{KL}}\left(q(z)||p(z)\right) + \log p(\boldsymbol{x}) - \mathbb{E}_{z\sim q(z)}\left[\log p(\boldsymbol{x}\mid z)\right]$$

$$\log p(\boldsymbol{x}) = \mathbb{E}_{z\sim q(z)}\left[\log p(\boldsymbol{x}\mid z)\right] - D_{\mathrm{KL}}\left(q(z)||p(z)\right) + \underbrace{D_{\mathrm{KL}}\left(q(z)||p(z\mid \boldsymbol{x})\right)}_{\geq 0}$$

$$\log p(\boldsymbol{x}) \geq \mathbb{E}_{z\sim q(z)}\left[\log p(\boldsymbol{x}\mid z)\right] - D_{\mathrm{KL}}\left(q(z)||p(z)\right) = \mathcal{L}(\boldsymbol{\theta_q})$$

- $\mathcal{L}(\boldsymbol{\theta_q})$= **Evidence Lower Bound (ELBO)**라고 부름

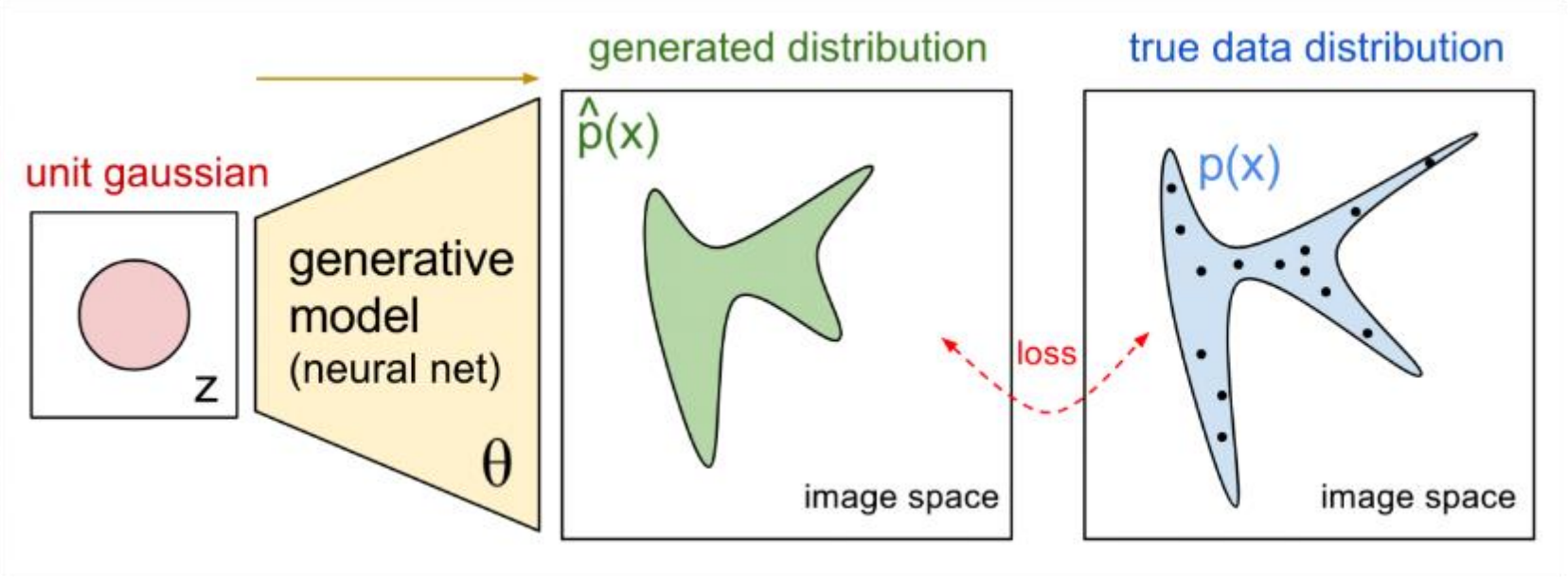
- $p(x|z)$ 에 추가적인 모델Parameter θ 가 붙으면 우리가 익숙한marginal log-likelihood의lower bound로볼수있다.

$$\log p(x;\theta) \geq \mathbb{E}_{z\sim q(z;\theta_q)}[\log p(x|z;\theta)] - D_{KL}(q(z;\theta_q)||p(z))$$
$$= \mathcal{L}(\theta_q,\theta) = \sum_z q(z;\theta_q)\,[\log \frac{p(x|z;\theta)p(z)}{q(z;\theta_q)}]$$

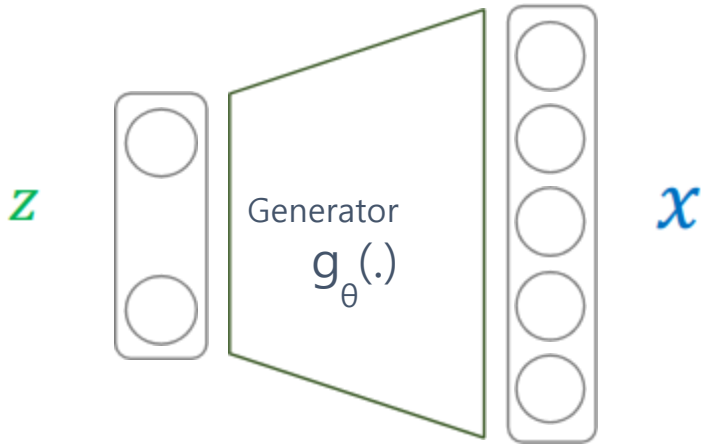
VAE

Variational Auto-Encoder

Deep Generative Model



Deep Generative Model



$z \sim p(z)$ Random variable

$g_\theta(\cdot)$ Deterministic function parameterized by θ

$x = g_\theta(z)$ Random variable

- Latent Variable z는target data x의특징을결정짓는다.
- z는또다른 latent code c와 concatination이될수있고, z 자체를 code로생각하여[0,…,9]을인풋으로 사용하는것도가능

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

- Generative Model의학습방법은data x에대한likelihood를최대화하는것이다. **MLE**

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

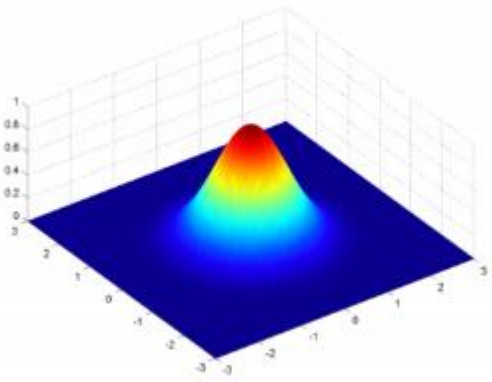
Probability density

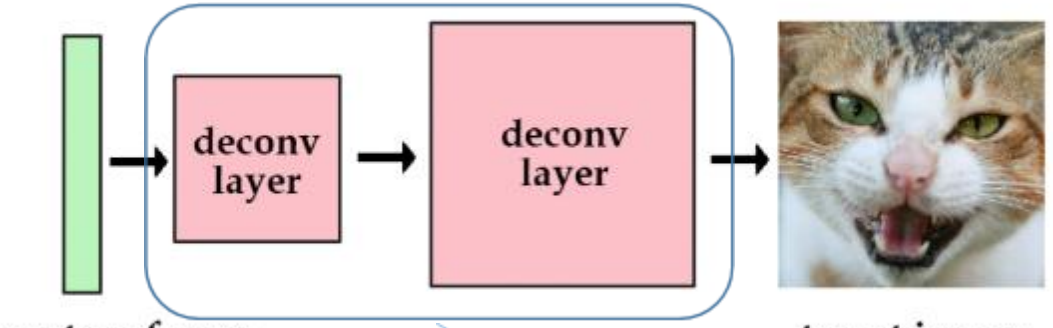
VAE 목표

디코더

(영상)데이터를 잘 생성하는확률모형 $P(X|Z;\theta)$ 학습하는것임

~~z~~~P(Z)는Gaussian 분포(*N*)사용.



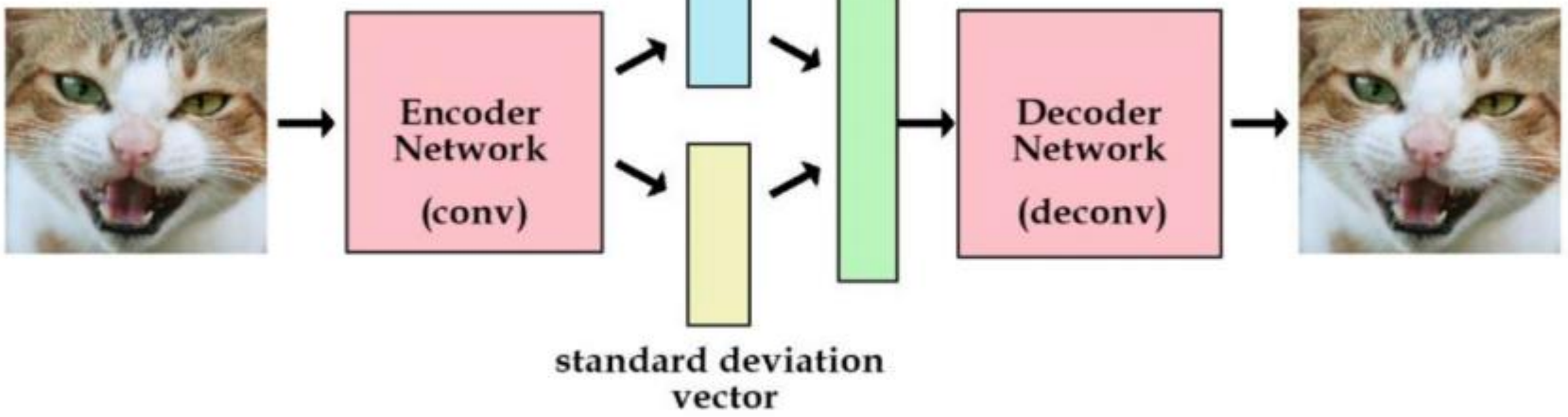


$$P(z) = \mathcal{N}(z|0, I)$$

가정: prior은 가우시안

$$P(X|z;\theta) = \mathcal{N}(X|f(z;\theta), \sigma^2 * I)$$

$q(z|x)$ $p(x|z)$



$$l_i(\theta, \phi) = -E_{z \sim q_\theta(z|x_i)}[\log p_\phi(x_i|z)] + KL(q_\theta(z|x_i)||p(z))$$

Reconstruction Loss

$\|x - x'\|$

KL Divergence
Regularizer

모형(VAE)을 어떻게 학습 시킴?

• **MLE**을통해서Generative Model의Parameter를 추정하려면
Marginal-log-likelihood $\log P(X;\theta)$ 를 maxima하면되겠다!?

디코더

$$\log p(x;\theta) = \log \sum_z p(x|z;f_\theta(z),\sigma^2 * I)p(z)$$

$P(X|Z;\theta)$

• 보통 Generative Model (e.g. GMM)에서도marginal log-Likelihood를
직접 최대화하기어려움ㅠㅠ 계산불가, No Analytic

• VAE는최적화대상인parameter (θ 가 확률모형 p(x|z) 속 $f_\theta(z)$,즉,
neural network의parameter θ 임으로 analytical 하게 최적의
Parameter를추정하는것은 어려움ㅠㅠ (문제발생)

⇒그럼
ELBO(VI=베이지안 접근)로
추정해볼까?

Posterior Likelihood Prior

$$P(\theta|x) = \frac{P(x|\theta)P(\theta)}{\int d\theta P(x|\theta)P(\theta)}$$

사후 정보 = 데이터로 얻어진 정보 + 사전 정보

ELBO and Variational Inference

VI를 통해서 임의의 z분포 $q(z;\theta_q)$ 를 만들고,

log-likelihood의 **Lower Bound**를 정할 수 있음.

q 끼겨넣기

$$\log p(x;\theta) = \log \sum_z \left[q(z;\theta_q) \frac{p(x|z;\theta)p(z)}{q(z;\theta_q)} \right]$$

목적: log-likelihood
Theta 학습

$$\geq \sum_z q(z;\theta_q) \left[\log \frac{p(x|z;\theta)p(z)}{q(z;\theta_q)} \right]$$

By Jensen's inequality

$$= \mathbb{E}_{z\sim q(z;\theta_q)}[\log p(x|z;\theta)] - D_{KL}(q(z;\theta_q)||p(z)) = \mathcal{L}(\boldsymbol{\theta_q}, \boldsymbol{\theta})$$

ELBO

• 계산이 쉬운 $\mathcal{L}(\boldsymbol{\theta_q}, \boldsymbol{\theta})$ 를 최대화하면!

원래 알고싶었던 이놈(Posterior)
Latent effective knowledge set.

KL divergence term이줄어들때 까지

$q(z;\theta_q)$ 는 알고싶었던 posterior $P(z|x)$ 를 닮아가며 학습 됨 $q^*(z)$ 를 얻음

$\log p(x;\theta)$ 는 최대화되도록학습이가능하다. ▪ **결론:** $p(z|x)$ 를 잘 근사한 posterior

VAE에서 일반적인 **VI**와 달리 **Amortize trick** 사용하는 이유

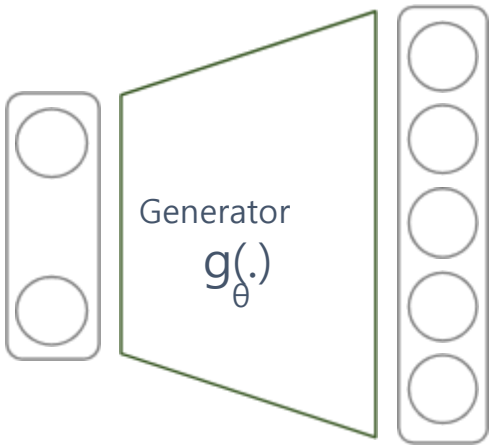
- 기존 Variational Inference에는 $q(z;\theta_q)$를 Gaussian으로 정했다.
- 그러나 x의 차원이 클때는, $q(z;\theta_q)$가 Posterior $P(z|x)$에 수렴하는 속도가 아주 느리다.
- 왜? $q(z;\theta_q)$ 가 Gaussian 인것은, 너무 단순해서 z와 x의 복잡한 관계인 $P(z|x)$를 근사하는 것이 어렵다.
- $q(z;\theta_q)$ 가 Posterior에 수렴하지 않는다면, Lower Bound가 tight해 지지 않는다. -> $\log P(x;\theta)$의 최대화도 어렵다.

Amortize trick

- $q(z;\theta_q)$ 가 너무 단순할 때 생기는 문제점을 해결하기 위해서 Amortized Variational Inference 사용한다.
- 말은 어려워 보이지만.. $q(z;\theta_q)$의 Parameter θ_q를 x에 대한 함수로 만드는 것이다.
- 즉, $q(z) = N(\mu_q, \sigma_q)$ 을 $q(z|x) = N(\mu(x), \sigma(x))$로 바꾼다.
- 이때, 복잡한 x와 z의 관계를 고려해서, $\mu(x)$, $\sigma(x)$를 또 다른 Neural Network로 정하였다.

VAE의 최종 목적 함수

$p(z|x) \approx q_\phi(z|x) \sim z$



Latent Variable Generated data

Optimization Problem 1 on ϕ: Variational Inference

$$\log(p(x)) \geq \mathbb{E}_{q_\phi(z|x)}[\log(p(x|z))] - KL(q_\phi(z|x)||p(z)) = ELBO(\phi)$$

Optimization Problem 2 on θ: Maximum likelihood

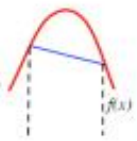
$$-\sum_i \log(p(x_i)) \leq -\sum_i \mathbb{E}_{q_\phi(z|x_i)}[\log(p(x_i|g_\theta(z)))] - KL(q_\phi(z|x_i)||p(z))$$

Final Optimization Problem

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

ELBO: Evidence Lower Bound in VAE

$$\log(p(x)) = \log\left(\int p(x|z)p(z)dz\right) \geq \int \log(p(x|z))p(z)dz$$



[Jensen's Inequality]

For concave functions f(.) f(E[x])≥E[f(x)]

$$\log(p(x)) = \log\left(\int p(x|z)\frac{p(z)}{q_\phi(z|x)}q_\phi(z|x)dz\right) \geq \int \log\left(p(x|z)\frac{p(z)}{q_\phi(z|x)}\right)q_\phi(z|x)dz$$

$$\log(p(x)) \geq \int \log(p(y|z))q_\psi(z|x)dz - \int \log\left(\frac{q_\phi(z|x)}{p(z)}\right)q_\psi(z|x)dz$$

$$= \mathbb{E}_{q_\phi(z|x)}[\log(p(x|z))] - KL(q_\phi(z|x)||p(z))$$

$ELBO(\phi)$ Variational lower bound
Evidence lower bound (ELBO)

x에 대하여 잘 모사
&& z에 대하여 잘 모사

ELBO를 최대화하는ϕ값을 찾는 다는 것은 $q_\phi(z|x)$와 $p(z|x)$간의 KL divergence를 줄이는 것과 같은 의미

ELBO: Evidence Lower Bound in VAE

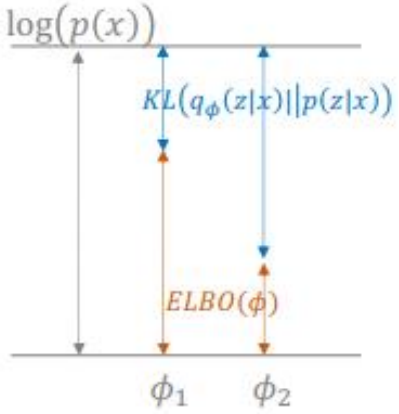
$$\log(p(x)) = \int \log(p(x)) q_\phi(z|x) dz \quad \leftarrow \int q_\phi(z|x)dz = 1$$

$$= \int \log\left(\frac{p(x,z)}{p(z|x)}\right) q_\phi(z|x) dz \quad \leftarrow p(x) = \frac{p(x,z)}{p(z|x)}$$

$$= \int \log\left(\frac{p(x,z)}{q_\phi(z|x)} \cdot \frac{q_\phi(z|x)}{p(z|x)}\right) q_\phi(z|x) dz$$

$$= \underbrace{\int \log\left(\frac{p(x,z)}{q_\phi(z|x)}\right) q_\phi(z|x) dz}_{ELBO(\phi)} + \underbrace{\int \log\left(\frac{q_\phi(z|x)}{p(z|x)}\right) q_\phi(z|x) dz}_{KL\left(q_\phi(z|x) \,\|\, p(z|x)\right)}$$

두 확률분포 간의 거리≥0



ELBO를 최대화하는ϕ값을 찾는 다는 것은 $q_\phi(z|x)$와 $p(z|x)$간의 KL divergence를 줄이는 것과 같은 의미

Variational Inference Objective and VAE

$$\log p(x) - D_{KL}(q_\theta(z|x)||p(z|x)) = \mathbb{E}_{z\sim q_\theta(z|x)}[\log\ p_\phi(x|z)] - D_{KL}(q_\theta(z|x)||p(z))$$

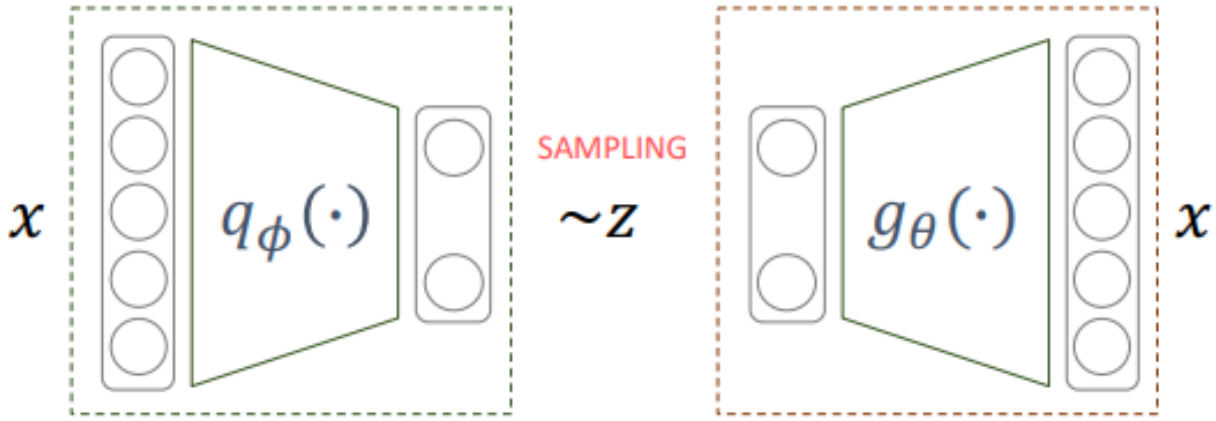
Reconstruction term — Regularizing Term

$$\log p(x) \geq \mathbb{E}_{z\sim q_\theta(z|x)}[\log\ p_\phi(x|z)] - D_{KL}(q_\theta(z|x)||p(z))$$

x에 대하여 잘 모사
&& z에 대하여 잘 모사

Neural Network의 관점

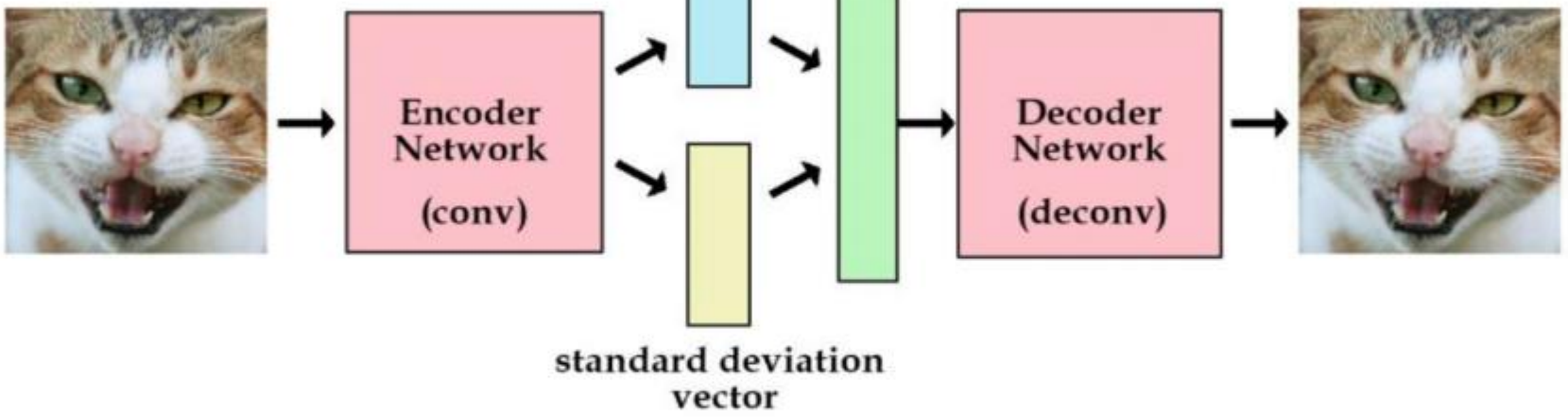
$$\arg\min_{\phi,\theta} \sum_i -\mathbb{E}_{q_\phi(z|x_i)}\left[\log\left(p(x_i|g_\theta(z))\right)\right] + KL\left(q_\phi(z|x_i)||p(z)\right)$$



Encoder
Posterior
Inference Network

Decoder
Generator
Generation Network

$q(z|x)$ $p(x|z)$



$$l_i(\theta, \phi) = -E_{z \sim q_\theta(z|x_i)}[\log p_\phi(x_i|z)] + KL(q_\theta(z|x_i)||p(z))$$

Reconstruction Loss

$\|x - x'\|$

KL Divergence
Regularizer

VAE Loss 해석 (1)

Variational inference를 위한 approximation class 중 선택

원 데이터에 대한 likelihood 선택

다루기 쉬운 확률 분포 중 선택

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

Reconstruction Error

- 현재 샘플된 z에 대한 negative log likelihood
- x_i에 대한 복원 오차 (AutoEncoder 관점)

Regularization

- 현재 샘플된 z에 대한 대한 추가 조건
- 샘플링되는 z들에 대한 통제성을 prior를 통해 부여, Variational distribution q(z|x)가 p(z)와 유사해야 한다는 조건을 부여

VAE Loss 해석 (2)

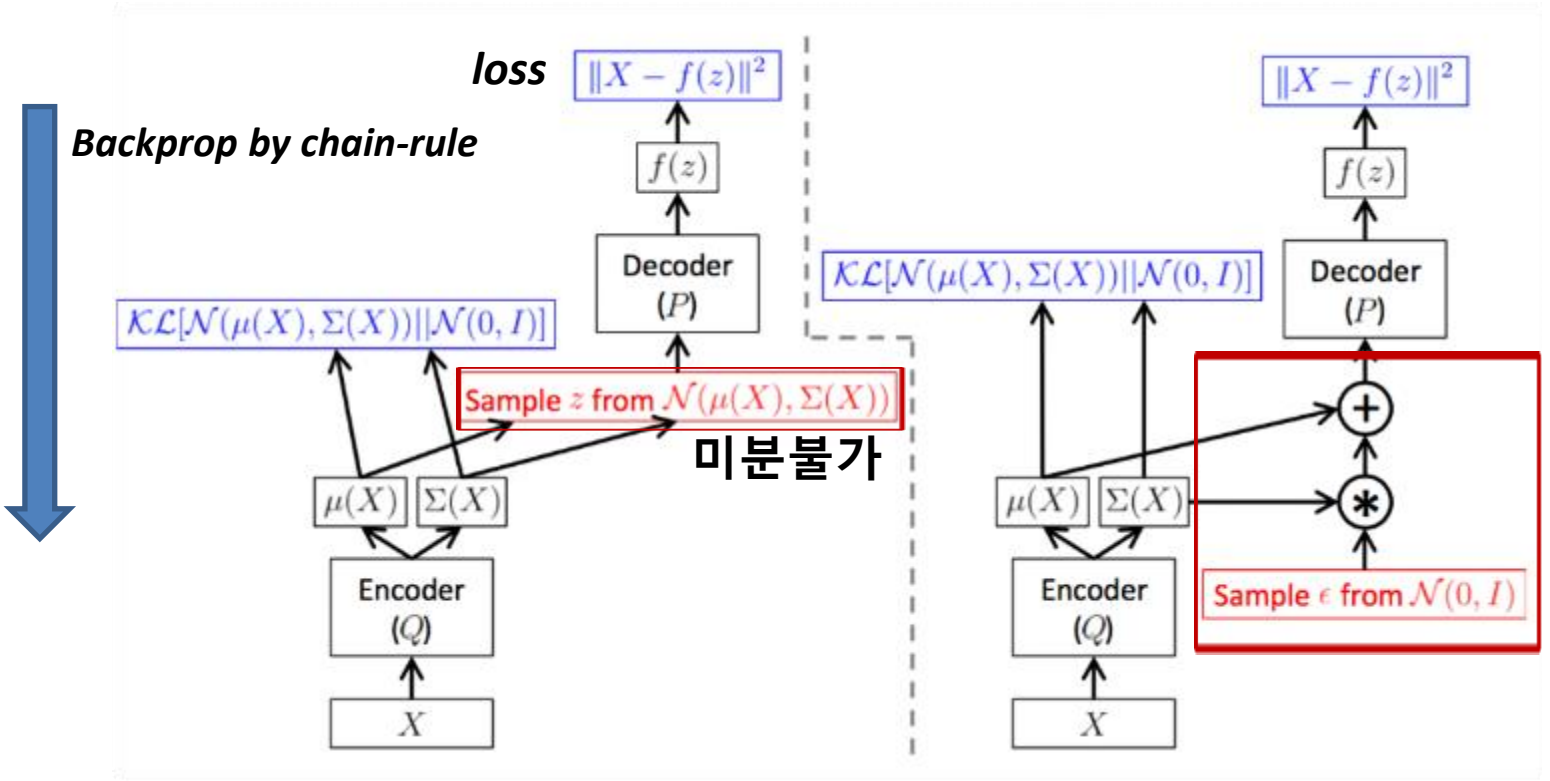
$$L_i(\phi,\theta,x_i) = -\mathbb{E}_{q_\phi(z|x_i)}\left[\log\left(p_\theta(x_i|z)\right)\right] + KL\left(q_\phi(z|x_i)||p(z)\right)$$

$$= \underbrace{-\mathbb{E}_{q_\phi(z|x_i)}\left[\log\left(p_\theta(x_i|z)\right)\right]}_{\text{Reconstruction Error}} \underbrace{- H\left(q_\phi(z|x_i)\right)}_{\text{Posterior Entropy}} + \underbrace{H\left(q_\phi(z|x_i), p(z)\right)}_{\text{Cross Entropy}}$$

Posterior에서 샘플링 된 z는 최대한 다양해야 한다
(mode collapse 방지 효과)

Posterior와 Prior의 정보량은 유사해야 한다

Re-parameterization Trick ?



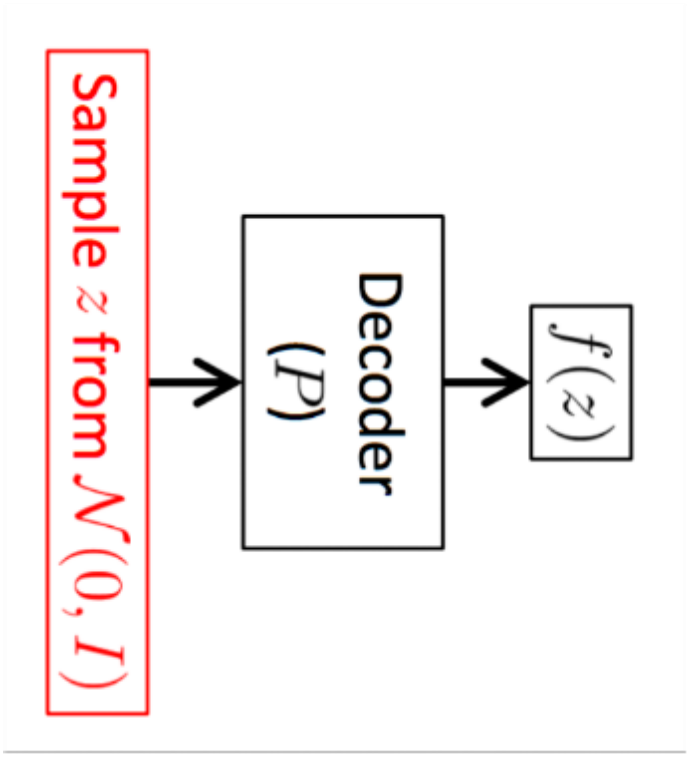
Sampling process $z^{i,l} \sim N(\mu_i, \sigma_i^2 I)$ ⇨ $z^{i,l} = \mu_i + \sigma_i^2 \odot \epsilon$, $\epsilon \sim N(0, I)$

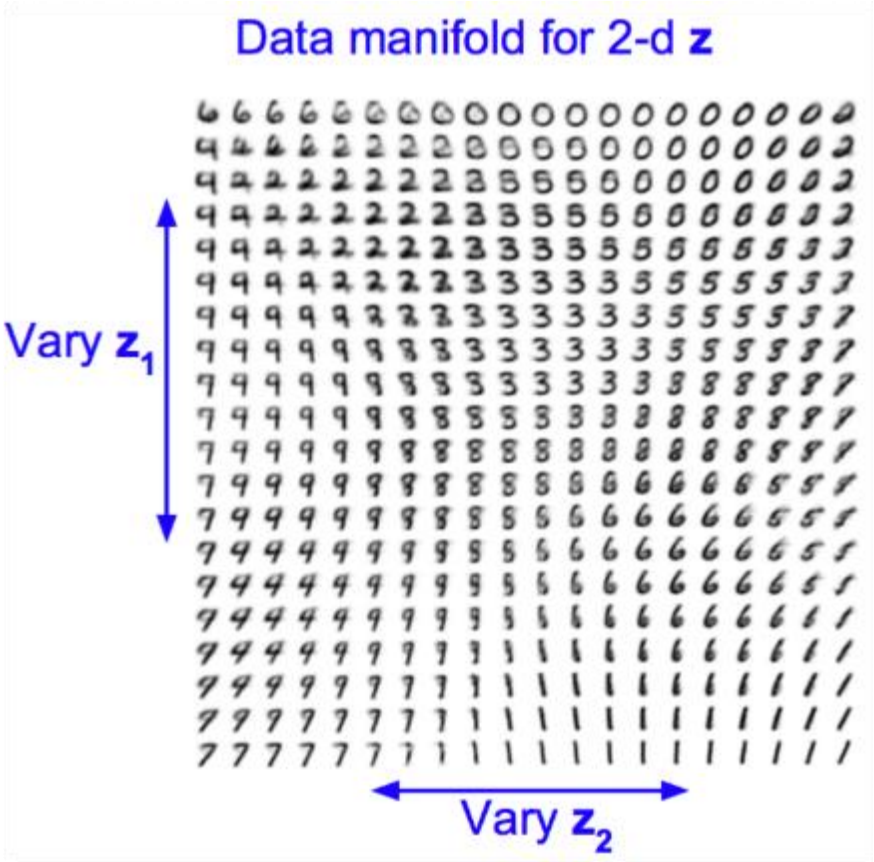
같은 파이프라인/결과 이지만 Backprop을 가능하게 함

https://home.zhaw.ch/~dueo/bbs/files/vae.pdf

Generator

- Data를 생성할 때는 z를 Decoder에 넣으면 된다.





MNIST Result

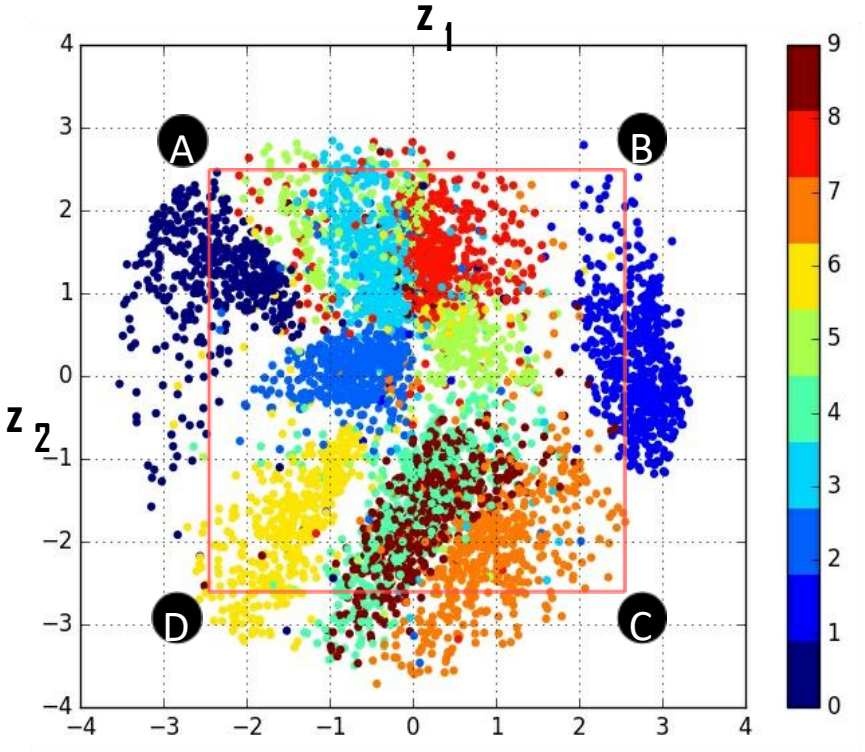
Input image

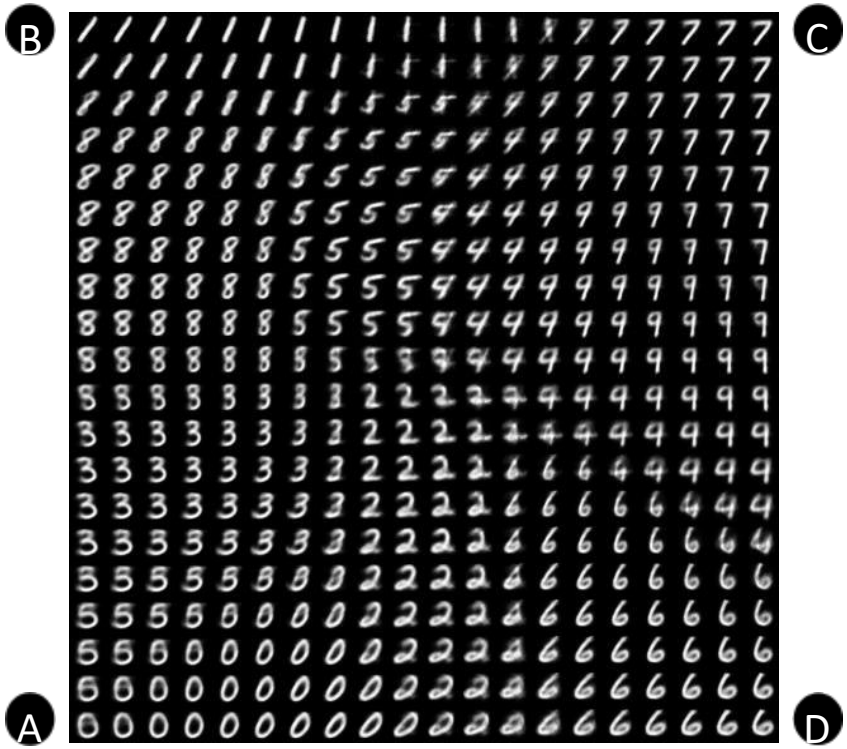
J = |z| =2

J = |z| =5

J = |z| =20

Learned Manifold (of latent space z)

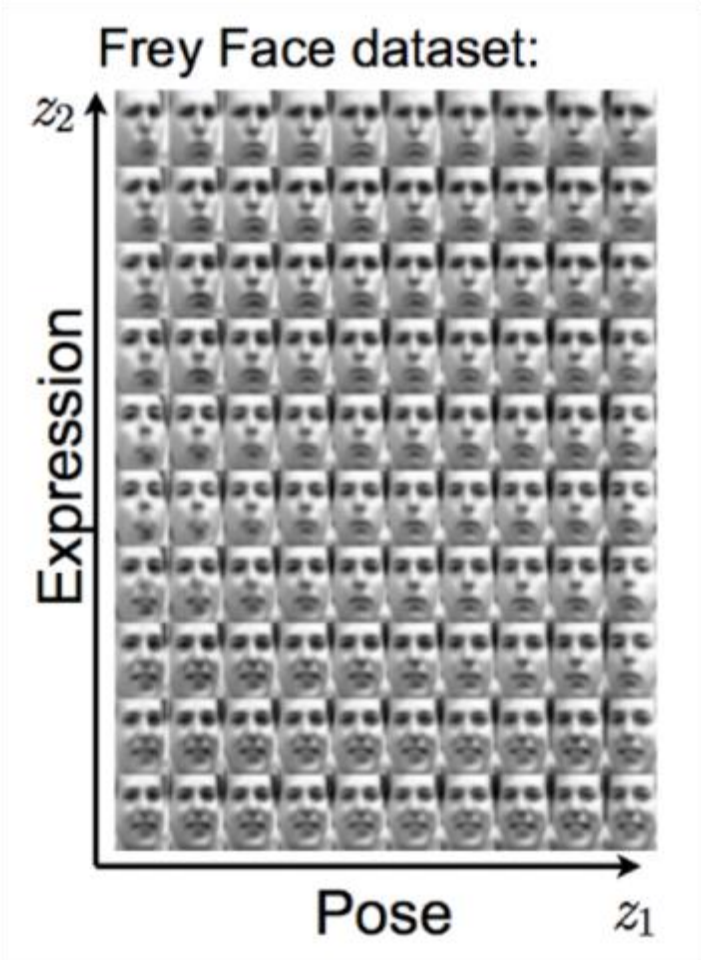




학습이잘되었을수록2D공간에서같은숫자들을생성하는z들은뭉쳐있고,
다른숫자들은생성하는z들은떨어져있어야한다.

https://github.com/hwalsuklee/tensorflow-mnist-VAE

Face Data 생성예시



Kingma and Welling. "*Auto-Encoding Variational Bayes, International Conference on Learning Representations.*" *ICLR*, 2014.
arXiv:1312.6114

VAE 장단점

- 학습이안정적인편
- 영상Generation뿐만아니라,주어진영상에서latent space의 posterior q(z|x)도함께학습시킨다. -> feature learning
- 평가기준이명확하다 Reconstruction Error, lower bound (estimate likelihood of data)

- GAN에 비해서 출력이선명하지않고 평균값 형태로 표시되는문제

  Reconstruction Loss -> Perceptual Loss
  Complex Posterior Inference -> Normalizing Flow
- Re-parameterization Trick이 Continuous한 함수에대해서만적용되는 문제

  Gumbel Categorical Distribution

Thanks.