VAE

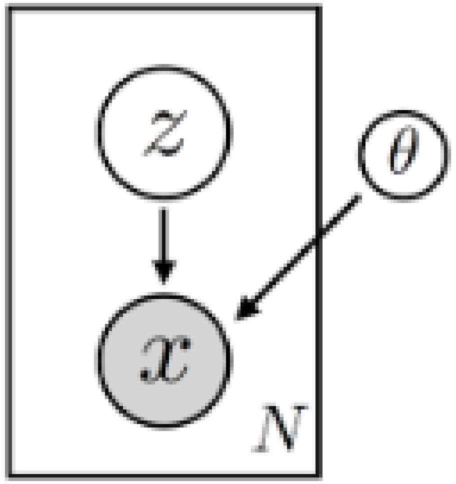
송용호

PDF 정의

$$p_\theta(x) = \int p_\theta(z) p_\theta(x|z) dz$$

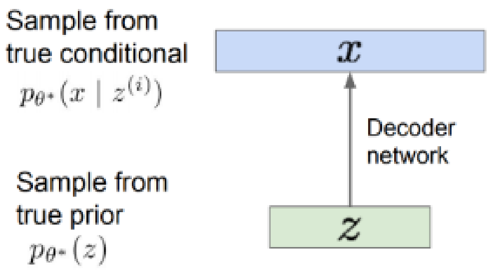
X는 iid, 동일 분포 따른다고 가정
세타는 파라미터
세타가 주어졌을 때, x가 나올 확률
이걸 최대화 하는게 목표.
Latent 변수 Z를 사용해서 나타내면 위 식과 같음.
이거 미분해서 SGD 할거임.

Latent variable



근데 z에서 샘플링하는거잖
분포는 가우시안이나 베르누이로 설정(다루기 쉬워서)
근데 모든 z에 대해 적분해야해 -> 못해
근데 MLE 문제를 풀라면 미분 해야해
이걸론 못함.

문제 2개



We want to estimate the true parameters θ^* of this generative model.

How to train the model?

Remember strategy for training generative models from FVBNs. Learn model parameters to maximize likelihood of training data

$$p_\theta(x) = \int p_\theta(z) p_\theta(x|z) dz$$

Q: What is the problem with this?

Intractable!

Kingma and Welling, "Auto-Encoding Variational Bayes", ICLR 2014

1. Z변수를 어떻게 구성?

가우시안 분포로부터 샘플링함
디코더는 z로 부터 데이터 생성

여러층의 레이어가 있으면

앞층은 Nomal을 latent로 바꿔주
고
뒤에는 이거 가지고 sampling 할
수 있는 pdf를 만들어내는것.

2. 적분을 어떻게 해결?

이거 다 계산 하 말고 Monte-Carlo estimatio로 하자.

근데 대부분 0값 가지니까, 많이 해야해.

가뜩이나 딥러닝 오래 걸리는데 많이 할 여유 없음

데이터에 종속적으로 z를 샘플링하자.(2)

$$p_\theta(x) = \int p_\theta(z) p_\theta(x|z) dz \approx \frac{1}{N} \sum_{i=1}^{N} p_\theta(x|z^{(i)}) \quad (2)$$

새로운 함수 정의

$$p_\theta(x) = \int p_\theta(z) p_\theta(x|z) dz \approx \frac{1}{N} \sum_{i=1}^{N} p_\theta(x|z^{(i)}) \quad (2)$$

Posterior density also intractable: $p_\theta(z|x) = p_\theta(x|z)p_\theta(z)/p_\theta(x)$

Intractable data likelihood

그래서 z given x 를 생각해보자 이거야.

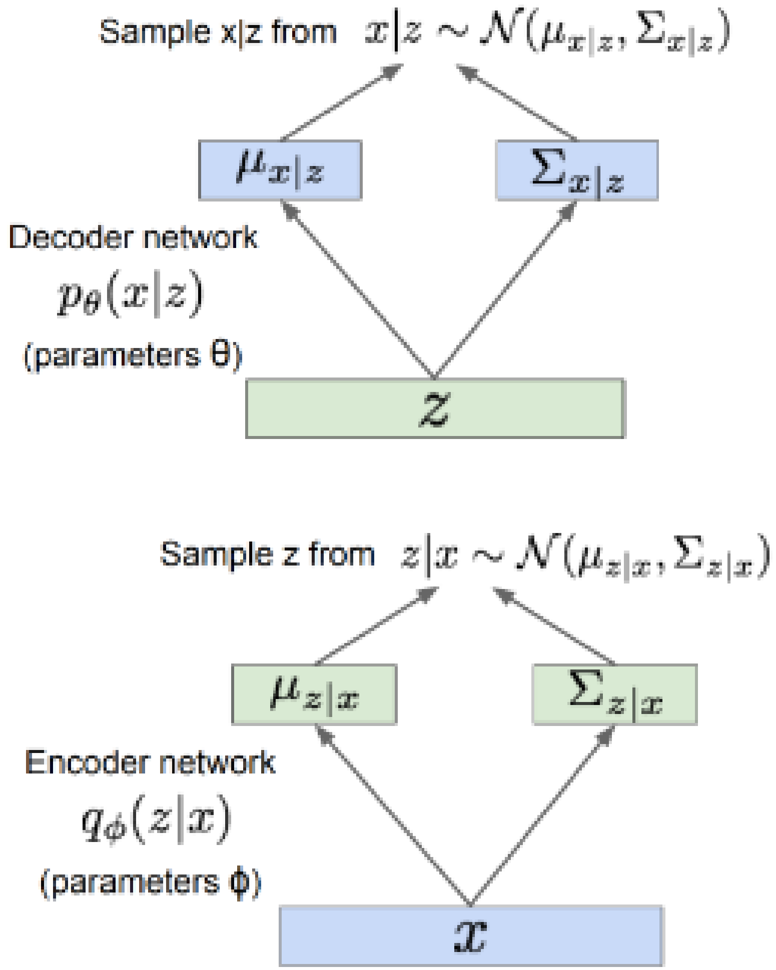
근데 X는 추적 불가능하잖

그니까 저 빨간 저것도 계산이 안돼

그래서 approximate하는 새로운 함수를 정의

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

새로운 함수 정의2



$q_\phi(z|x)$ 요건 원래 posterior를 근사했기 때문에 오류 존재

따라서 Object functio에 lower bound 정의 할거임.

그 전에 VAE 구조부터 봐보자.

1. X 받아서 latent에 분포 만듬

2. 가우시안이라 가정하고, data dependent한 가우시안 분포로부터 z 샘플링

3. 샘플링된 z가지고 다시 가정한 분포로 디코더가 분포를 뱉어냄

4. X를 이 분포로부터 샘플링.

ELBO(Evidence Lower Bound)

https://velog.io/@thdfydgh/ELBO

$$\log p_\theta(x^{(i)}) = \mathbf{E}_{z\sim q_\phi(z|x^{(i)})}\left[\log p_\theta(x^{(i)})\right] \quad (p_\theta(x^{(i)}) \text{ Does not depend on } z)$$

$$= \mathbf{E}_z\left[\log \frac{p_\theta(x^{(i)} \mid z)p_\theta(z)}{p_\theta(z \mid x^{(i)})}\right] \quad (\text{Bayes' Rule})$$

$$= \mathbf{E}_z\left[\log \frac{p_\theta(x^{(i)} \mid z)p_\theta(z)}{p_\theta(z \mid x^{(i)})}\frac{q_\phi(z \mid x^{(i)})}{q_\phi(z \mid x^{(i)})}\right] \quad (\text{Multiply by constant})$$

$$= \mathbf{E}_z\left[\log p_\theta(x^{(i)} \mid z)\right] - \mathbf{E}_z\left[\log \frac{q_\phi(z \mid x^{(i)})}{p_\theta(z)}\right] + \mathbf{E}_z\left[\log \frac{q_\phi(z \mid x^{(i)})}{p_\theta(z \mid x^{(i)})}\right] \quad (\text{Logarithms})$$

$$= \underbrace{\mathbf{E}_z\left[\log p_\theta(x^{(i)} \mid z)\right] - D_{KL}(q_\phi(z \mid x^{(i)}) \,\|\, p_\theta(z))}_{\mathcal{L}(x^{(i)},\theta,\phi)} + \underbrace{D_{KL}(q_\phi(z \mid x^{(i)}) \,\|\, p_\theta(z \mid x^{(i)}))}_{\geq 0}$$

Reconstruct the input data

Make approximate posterior distribution close to prior

$$\log p_\theta(x^{(i)}) \geq \mathcal{L}(x^{(i)},\theta,\phi)$$

Variational lower bound ("ELBO")

$$\theta^*, \phi^* = \arg\max_{\theta,\phi} \sum_{i=1}^{N} \mathcal{L}(x^{(i)},\theta,\phi)$$

Training: Maximize lower bound

의미 :
1항 : q의 x로부터 sample한 z가 있음.
Z가지고 p가 x를 생성할 likelihood

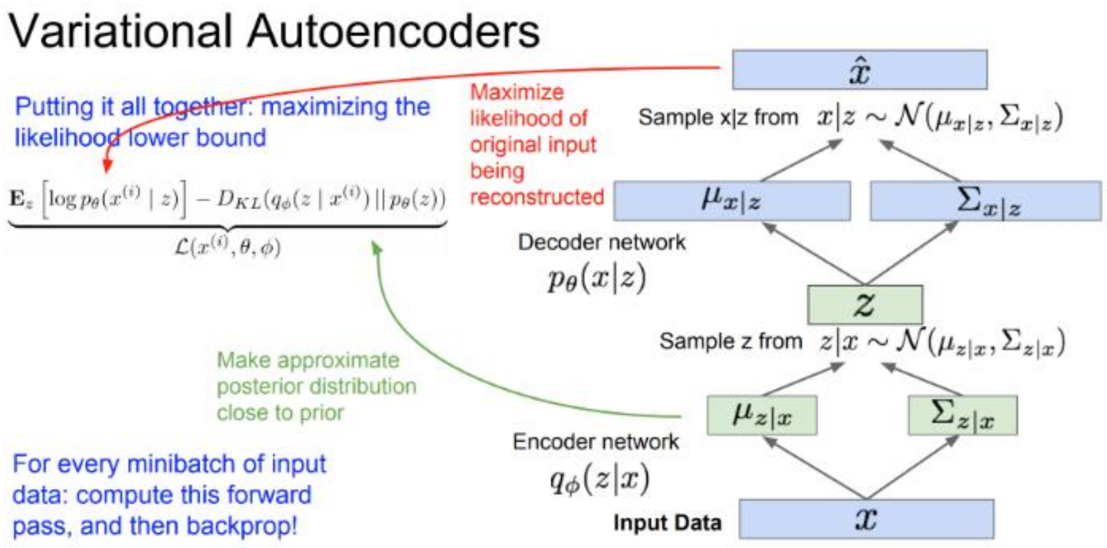
2항 : prior인 z와 approximate된 q 사이의 KLD

3항 : z given x는 추적 불가능, 상수취급

1,2항만 가지고 하한 설정 가능,

미분 가능해짐, MLE를 위해 backpropa 가능

train



X를 넣으면, latent space상의 mean, var 뱉음
: dim당 하나씩

그럼 mean과 var이 posterio를 나타내게 됨.

그럼 piror과 KLD를 구할 수 있어.

Z로부터 decoder는 data space상의 mean과 var을 뱉어

그럼 ELBO 1항(recon) 구할 수 있음

Backprop화면 VAE 학습 완료.

Monte carlo estimation

$$\mathcal{L}(x^{(i)}, \theta, \phi) = \mathbb{E}_z[\log p_\theta(x^{(i)}|z)] - D_{KL}(q_\phi(z|x^{(i)}) \| p_\theta(z))$$

근데 1항에 파이에 대해 미분하는건 문제가 있음

ϕ에 대해서 미분하는 것은 문제가 있습니다. 그 이유는 첫 번째 항인 $\mathbb{E}_{z\sim q_\phi(z|x)}[\log p_\theta(x|z)]$의 기댓값을 구하기 위해서는, $\mathbb{E}_{q_\phi(z|x)}$의 형태로 되어 있어야 합니다. 그러나 $\log p_\theta(x|z)$의 기댓값을 구하기 위해서는 z를 $q_\phi(z|x)$로부터 샘플링해야 합니다. 이러한 샘플링 과정에서는 그라디언트가 ϕ에 대해 직접적으로 전달되지 않습니다. 따라서 미분이 어려워집니다.

근데 1항에 파이에 대해 미분하는건 문제가 있음

$$\nabla_\phi \mathbb{E}_{q_\phi(z)}[f(z)] = \int \nabla_\phi q_\phi(z) f(z) dz$$

$$= \int q_\phi(z) \frac{\nabla_\phi q_\phi(z)}{q_\phi(z)} f(z) dz$$

$$= \mathbb{E}_{q_\phi(z)}[f(z) \nabla_\phi \log q_\phi(z)]$$

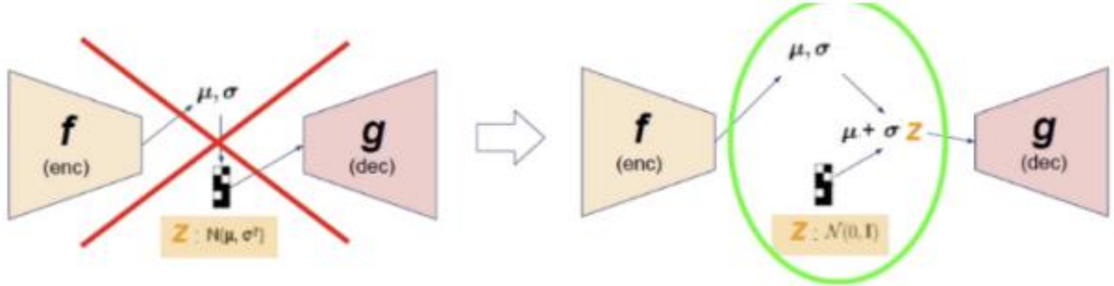
그래서 Monte carlo estimatio을 사용

Sample해서 평균 때리잔건데,

그냥 1번 뽑는걸로 퉁 침.(여기서 L=1)

$$\frac{1}{L}\sum_{l=1}^{L} f(z^{(l)}) \nabla_\phi \log q_\phi(z^{(l)})$$

Reparameterization trick



Sampling process $z^{i,l} \sim N(\mu_i, \sigma_i^2 I)$

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

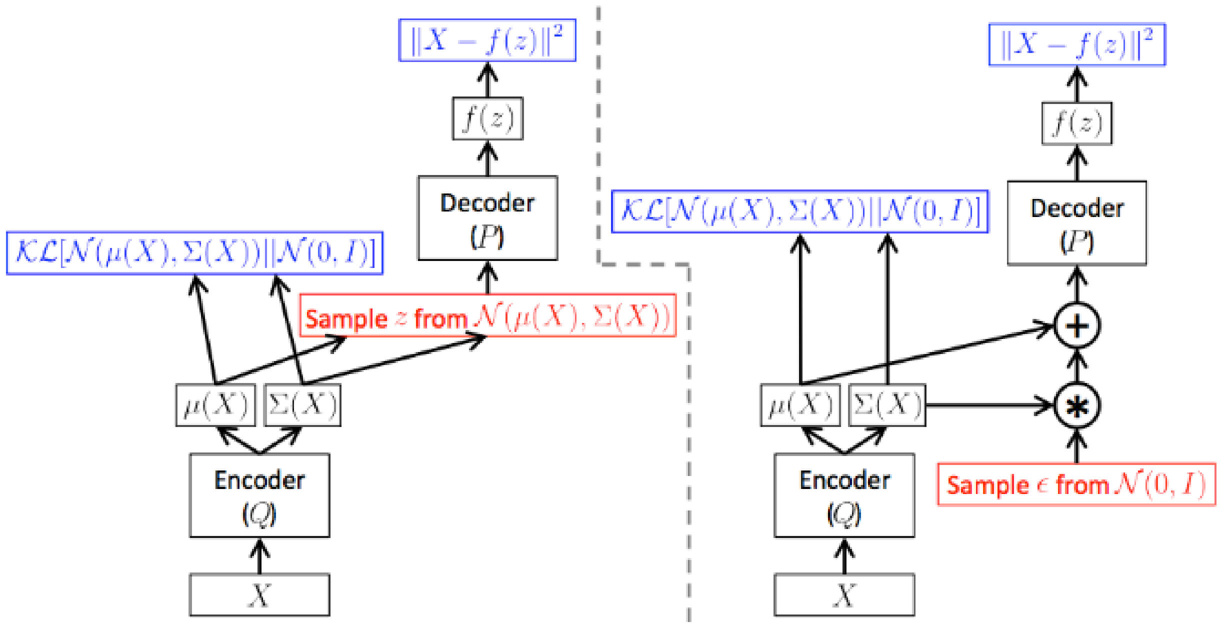
Same distribution!
But it makes backpropagation possible!

근데 몬테 카를로는 분산이 너무 커서 gradient 전달이 안됨.

그래서 reparm trick을 사용.

다들 알거라 생각하고 skip

Auto Encoding Variational Bayes



Algorithm 1 Minibatch version of the Auto-Encoding VB (AEVB) algorithm. Either of the two SGVB estimators in section 2.3 can be used. We use settings $M = 100$ and $L = 1$ in experiments.

$\boldsymbol{\theta}, \boldsymbol{\phi} \leftarrow$ Initialize parameters
repeat
  $\mathbf{X}^M \leftarrow$ Random minibatch of M datapoints (drawn from full dataset)
  $\boldsymbol{\epsilon} \leftarrow$ Random samples from noise distribution $p(\boldsymbol{\epsilon})$
  $\mathbf{g} \leftarrow \nabla_{\boldsymbol{\theta}, \boldsymbol{\phi}} \widetilde{\mathcal{L}}^M(\boldsymbol{\theta}, \boldsymbol{\phi}; \mathbf{X}^M, \boldsymbol{\epsilon})$ (Gradients of minibatch estimator (8))
  $\boldsymbol{\theta}, \boldsymbol{\phi} \leftarrow$ Update parameters using gradients $\mathbf{g}$ (e.g. SGD or Adagrad [DHS10])
until convergence of parameters $(\boldsymbol{\theta}, \boldsymbol{\phi})$
return $\boldsymbol{\theta}, \boldsymbol{\phi}$

reparameterization trick을 그림으로 보자면

encoder로부터 구한 data dependent한 mean과 variance를 가지고 posterior를 만듬

그 posterior로부터 z를 샘플링한 다음에 그 z를 가지고 decoder는 data를 generation

하지만 reparametization을 하면

computation graph 내의 sampling 과정이

noise sampling이 되어 옆으로 빠져

따라서 Back propagation을 통해 decoder output으로부터 encoder까지 gradient가 전달