Variational Auto encoder(VAE)

Auto-Encoding Variational Bayes

Diederik P. Kingma
Machine Learning Group
Universiteit van Amsterdam
dpkingma@gmail.com

Max Welling
Machine Learning Group
Universiteit van Amsterdam
welling.max@gmail.com

Abstract

How can we perform efficient inference and learning in directed probabilistic models, in the presence of continuous latent variables with intractable posterior

2020.01.06
박진수

Structure of Generative model

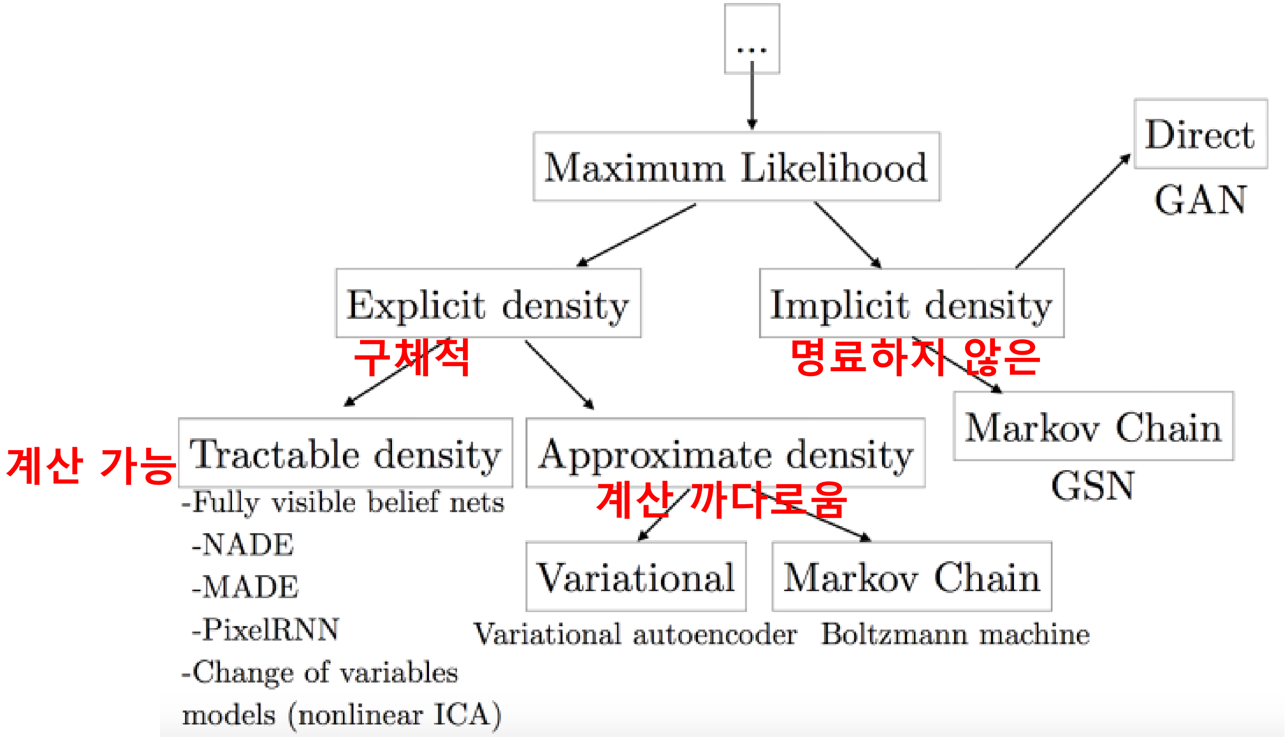
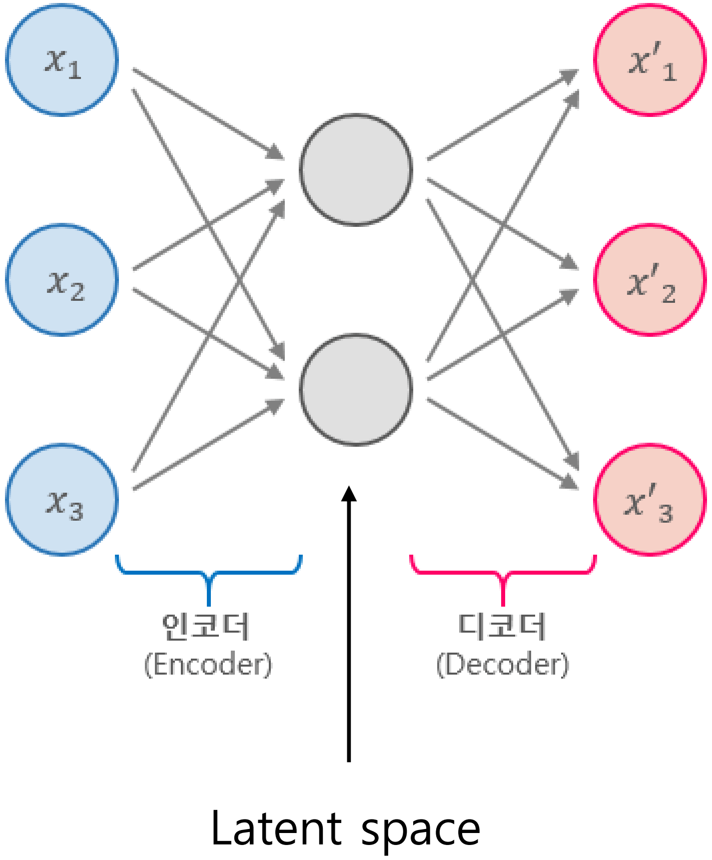


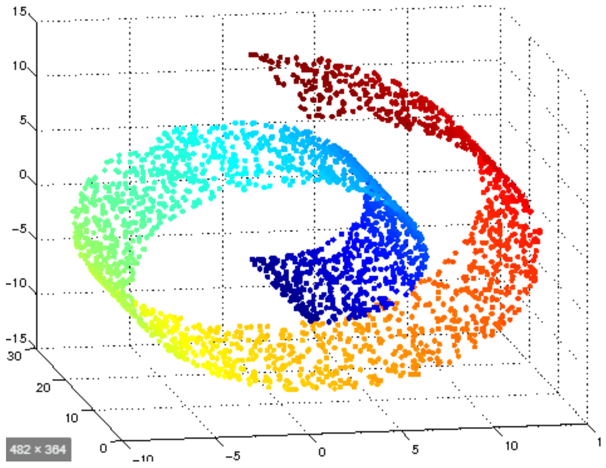
Figure is borrowed from Ian Goodfellow - NIPS 2016 Tutorial Generative Adversarial Networks

Auto encoder

Variational inference

Brief Intro. Autoencoder





1. Unsupervised learning -> like supervised learning
2. Manifold learning -> assume sub space
3. Dimensionality reduction -> important features
   (ex; PCA(linear activation))

Bayesian inference

<Bayes rule>

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Direct

- Calculate formula (using conjugate)

Indirect (Approximation)

- Sampling
- Variational inference

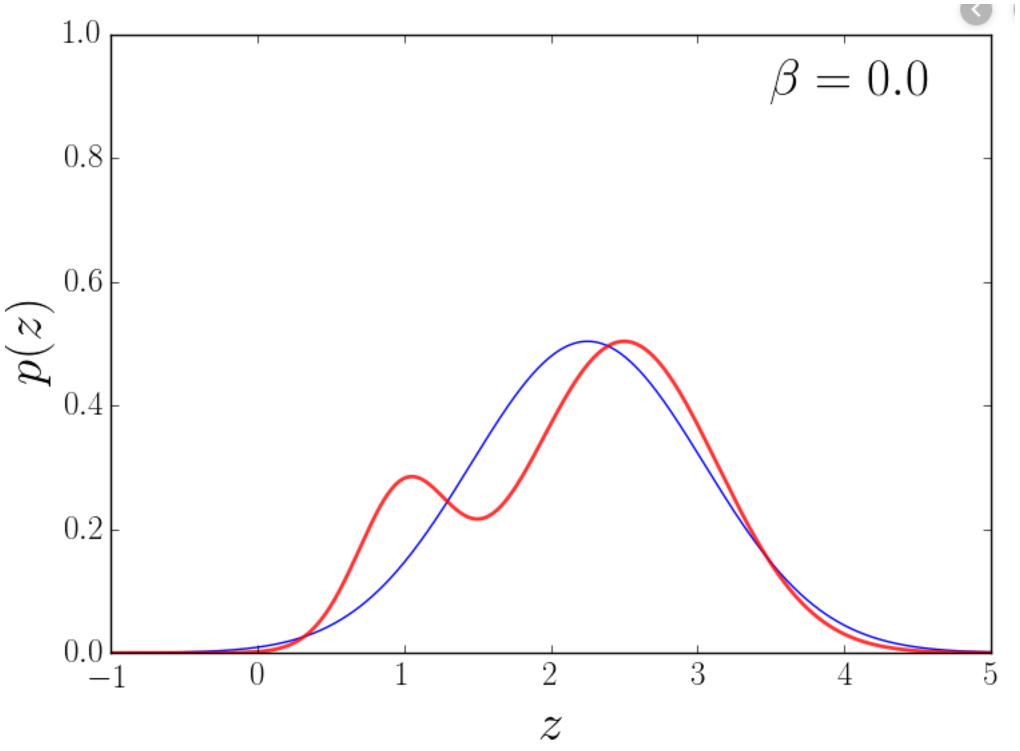
Variational Inference

$$P(\mathbf{Z} \mid \mathbf{X}) = \frac{P(\mathbf{X} \mid \mathbf{Z})P(\mathbf{Z})}{P(\mathbf{X})} = \frac{P(\mathbf{X} \mid \mathbf{Z})P(\mathbf{Z})}{\int_{\mathbf{Z}} P(\mathbf{X}, \mathbf{Z})\, d\mathbf{Z}}$$

$$P(\mathbf{Z} \mid \mathbf{X}) \approx Q(\mathbf{Z}).$$

$$D_{\mathrm{KL}}(Q \parallel P) \triangleq \sum_{\mathbf{Z}} Q(\mathbf{Z}) \log \frac{Q(\mathbf{Z})}{P(\mathbf{Z} \mid \mathbf{X})}.$$

- 단순 선형 방정식이라도 3개의 모수가 필요
; 기울기, 절편, 오차



Variational Auto Encoder

VAE = Generative model

- X가 생성될 수 있게 분포 p(X)에 대해서 학습한다면 어떨까?

Variational Auto Encoder

$\mathbf{x}$: data we want to model

$\mathbf{z}$: *latent variables*

$p(\mathbf{x})$: probability distribution of the data $\mathbf{x}$

$p(\mathbf{z})$: probability distribution of the latent variables $\mathbf{z}$

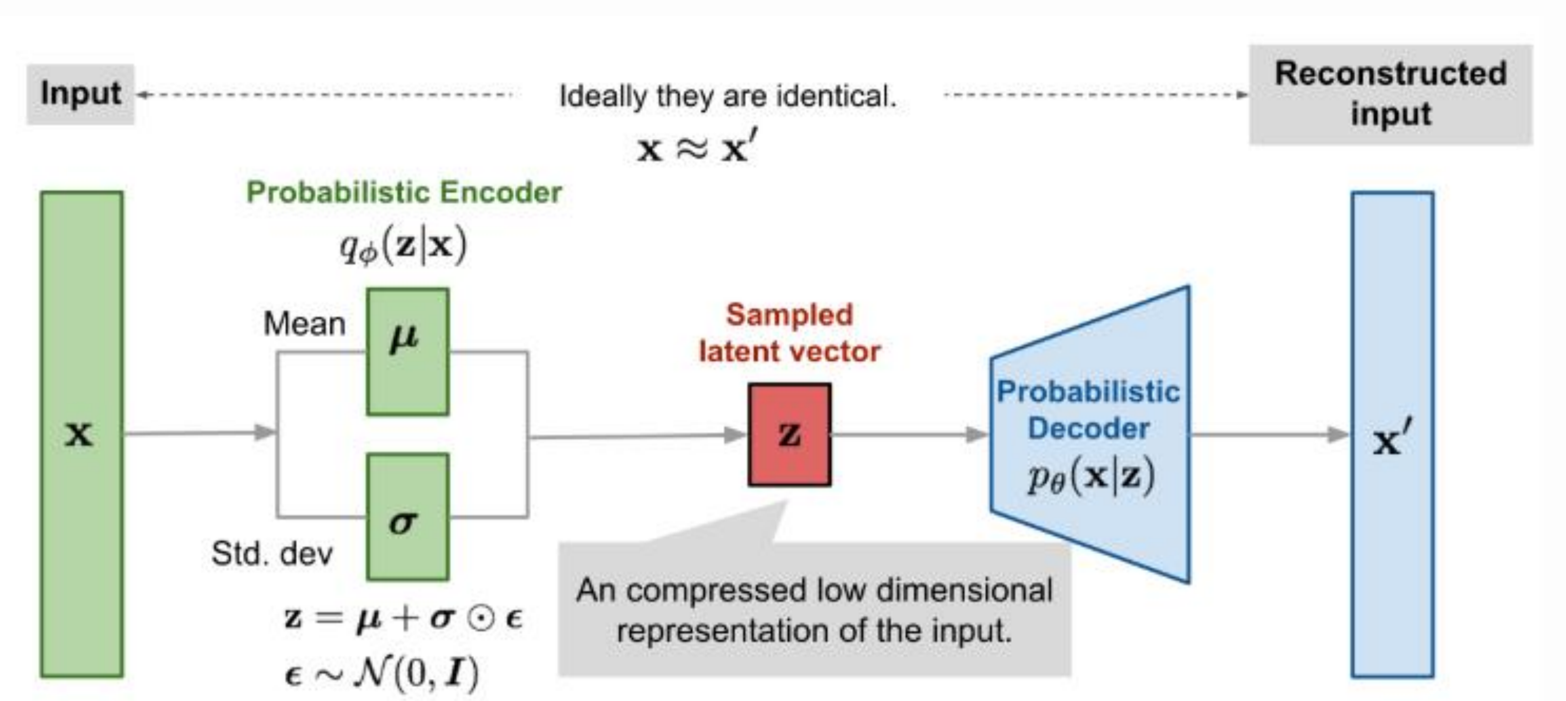
$p(\mathbf{x}|\mathbf{z})$: conditional probability distribution of the data given latent variables

$$p(\mathbf{x}) = \int p(\mathbf{x}, \mathbf{z}) \mathrm{d}\mathbf{z} = \int p(\mathbf{x}|\mathbf{z}) p(\mathbf{z}) \mathrm{d}\mathbf{z}$$

- 목표는 **P(X)**를 구하는 것이다.
- P(X)를 구하기 위해서는 모수Z를 구해야 한다.**(P(X|Z)**
  -> **P(Z)**도 알아야한다.
- P(Z)를 구하기 위해선 데이터에 대한 정보가 있어야 한다.**P(Z|X)**

P(X) -> P(Z|X) -> Z -> P(X|Z) -> P(X')

Variational Auto Encoder



P(X) P(Z|X) ≒ q(Z|X) Z P(X|Z) P(X)

우리가 prior로 선택할 p(z)는 최대한 직관적이고 간단한 것이 좋다.
-> 분석자의 통제력이 커진다.

Variational Auto Encoder

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

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

Solution: In addition to decoder network modeling $p_\theta(x|z)$, define additional encoder network $q_\phi(z|x)$ that approximates $p_\theta(z|x)$

Will see that this allows us to derive a lower bound on the data likelihood that is tractable, which we can optimize

Variational Auto Encoder

$$p(\mathbf{z}|\mathbf{x}) = \frac{p(\mathbf{x}|\mathbf{z})p(\mathbf{z})}{p(\mathbf{x})} \quad ≒ \quad q(z|x)$$

"복잡한 posterior를 q(z|x)로 근사"

Variational Auto Encoder

$$D_{KL}[q_\lambda(\mathbf{z}|\mathbf{x})||p(\mathbf{z}|\mathbf{x})] = \mathbb{E}_{q_\lambda}\left[\log \frac{q_\lambda(\mathbf{z}|\mathbf{x})}{p(\mathbf{z}|\mathbf{x})}\right] = \mathbb{E}_{q_\lambda}\left[\log q_\lambda(\mathbf{z}|\mathbf{x}) - \log p(\mathbf{z}|\mathbf{x})\right]$$

$$\begin{aligned} D_{KL}[q_\lambda(\mathbf{z}|\mathbf{x})||p(\mathbf{z}|\mathbf{x})] &= \mathbb{E}_{q_\lambda}\left[\log q_\lambda(\mathbf{z}|\mathbf{x}) - \log p(\mathbf{z}|\mathbf{x})\right] \\ &= \mathbb{E}_{q_\lambda}\left[\log q_\lambda(\mathbf{z}|\mathbf{x}) - \log \frac{p(\mathbf{x},\mathbf{z})}{p(\mathbf{x})}\right] \\ &= \mathbb{E}_{q_\lambda}\left[\log q_\lambda(\mathbf{z}|\mathbf{x}) - \log p(\mathbf{x},\mathbf{z}) + \log p(\mathbf{x})\right] \\ &= \mathbb{E}_{q_\lambda}\left[\log q_\lambda(\mathbf{z}|\mathbf{x}) - \log p(\mathbf{x},\mathbf{z})\right] + \log p(\mathbf{x}) \end{aligned}$$

"KL divergency가 최소화 되게 하는 q(z|x)를 선택해야 한다"

Variational Auto Encoder

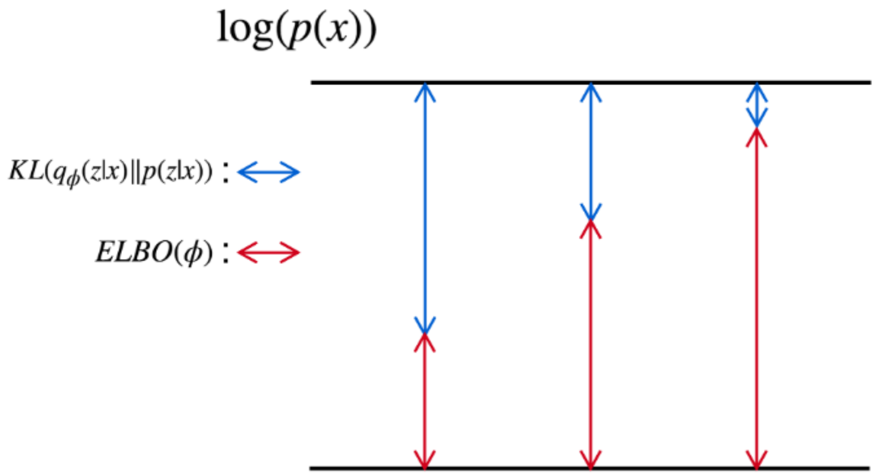
$$
\begin{aligned}
q_\lambda^*(\mathbf{z}|\mathbf{x}) &= \arg\min_\lambda D_{KL}(q_\lambda(\mathbf{z}|\mathbf{x})||p(\mathbf{z}|\mathbf{x})) \\
&= \arg\min_\lambda \mathbb{E}_q \left[\log q_\lambda(\mathbf{z}|\mathbf{x}) - \log p(\mathbf{x}, \mathbf{z})\right] + \log p(\mathbf{x})
\end{aligned}
$$

ELBO를 다음과 같이 가정하면,

$$
\mathrm{ELBO}(\lambda) = \mathbb{E}_{q_\lambda}\left[\log p(\mathbf{x}, \mathbf{z})\right] - \mathbb{E}_{q_\lambda}\left[\log q_\lambda(\mathbf{z}|\mathbf{x})\right]
$$

$$
\log p(\mathbf{x}) = \mathrm{ELBO}(\lambda) + D_{KL}(q_\lambda(\mathbf{z}|\mathbf{x})||p(\mathbf{z}|\mathbf{x}))
$$

Variational Auto Encoder



$$\log p(\mathbf{x}) = \mathrm{ELBO}(\lambda) + D_{KL}(q_\lambda(\mathbf{z}|\mathbf{x})||p(\mathbf{z}|\mathbf{x}))$$

$$\log p(\mathbf{x}) \geq \mathrm{ELBO}(\lambda) = \mathbb{E}_{q_\lambda}\left[\log p(\mathbf{x}, \mathbf{z})\right] - \mathbb{E}_{q_\lambda}\left[\log q_\lambda(\mathbf{z}|\mathbf{x})\right]$$

Always $D_{KL} \geq 0$

Minimizing KL divergence is equivalent maximizing ELBO

Variational Auto Encoder

$$
\begin{aligned}
\mathrm{ELBO}(\lambda) &= \mathbb{E}_{q_\lambda}\left[\log p(\mathbf{x},\mathbf{z})\right] - \mathbb{E}_{q_\lambda}\left[\log q_\lambda(\mathbf{z}|\mathbf{x})\right] \\
&= \mathbb{E}_{q_\lambda}\left[\log p(\mathbf{x},\mathbf{z}) - \log q_\lambda(\mathbf{z}|\mathbf{x})\right] \\
&= \mathbb{E}_{q_\lambda}\left[\log p(\mathbf{x}|\mathbf{z})p(\mathbf{z}) - \log q_\lambda(\mathbf{z}|\mathbf{x})\right] \\
&= \mathbb{E}_{q_\lambda}\left[\log p(\mathbf{x}|\mathbf{z}) + \log p(\mathbf{z}) - \log q_\lambda(\mathbf{z}|\mathbf{x})\right] \\
&= \mathbb{E}_{q_\lambda}\left[\log p(\mathbf{x}|\mathbf{z}) - (\log q_\lambda(\mathbf{z}|\mathbf{x}) - \log p(\mathbf{z}))\right] \\
&= \mathbb{E}_{q_\lambda}\left[\log p(\mathbf{x}|\mathbf{z})\right] - D_{KL}\left[q_\lambda(\mathbf{z}|\mathbf{x})||p(\mathbf{z})\right]
\end{aligned}
$$

Variational Auto Encoder

Loss : 1) Reconstruction loss + 2) regularization loss

```
generation_loss = mean(square(generated_image - real_image))
latent_loss = KL-Divergence(latent_variable, unit_gaussian)
loss = generation_loss + latent_loss
```

1) Reconstruction loss

$$\mathbb{E}_{q_\lambda(\mathbf{z}|\mathbf{x})}[\log p(\mathbf{x}|\mathbf{z})]$$

- Latent variable(z)가 주어졌을 때 디코더가 얼마나 효과적으로 데이터(x)를 잘 recon 하는지
- loss function L(x−x^) ; x(input), x^(recon)

2) Regularization loss

$$D_{KL}[q_\lambda(\mathbf{z}|\mathbf{x})||p(\mathbf{z})]$$

- 인코더에서 근사한 함수 q와 다루기 쉬운 함수로 우리가 선택한 p와의 거리
- 생성한 변수들을 잘 다루기 위해 P는 단순하면 좋다.

Variational Auto Encoder

***Reparameterization trick*

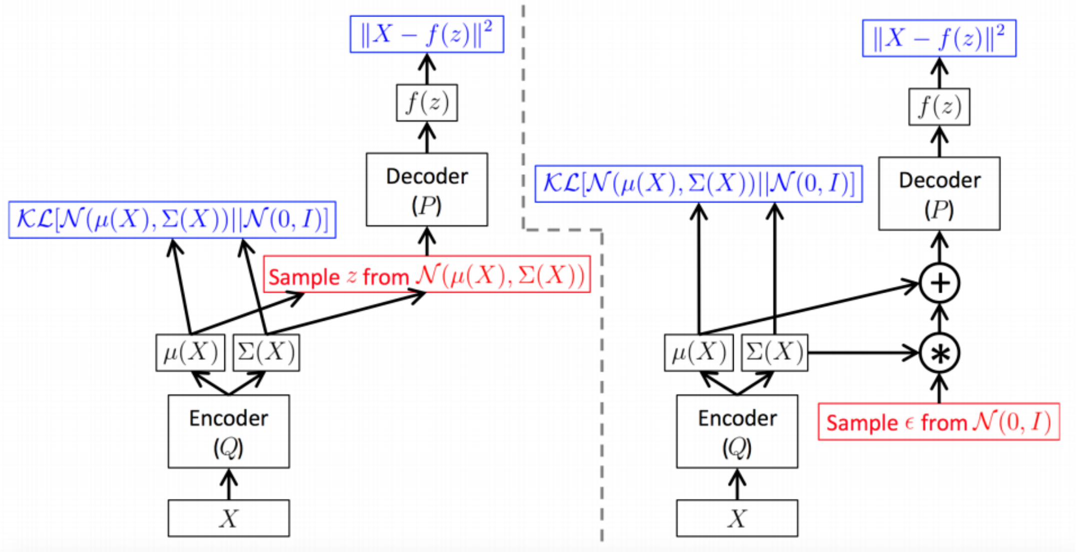
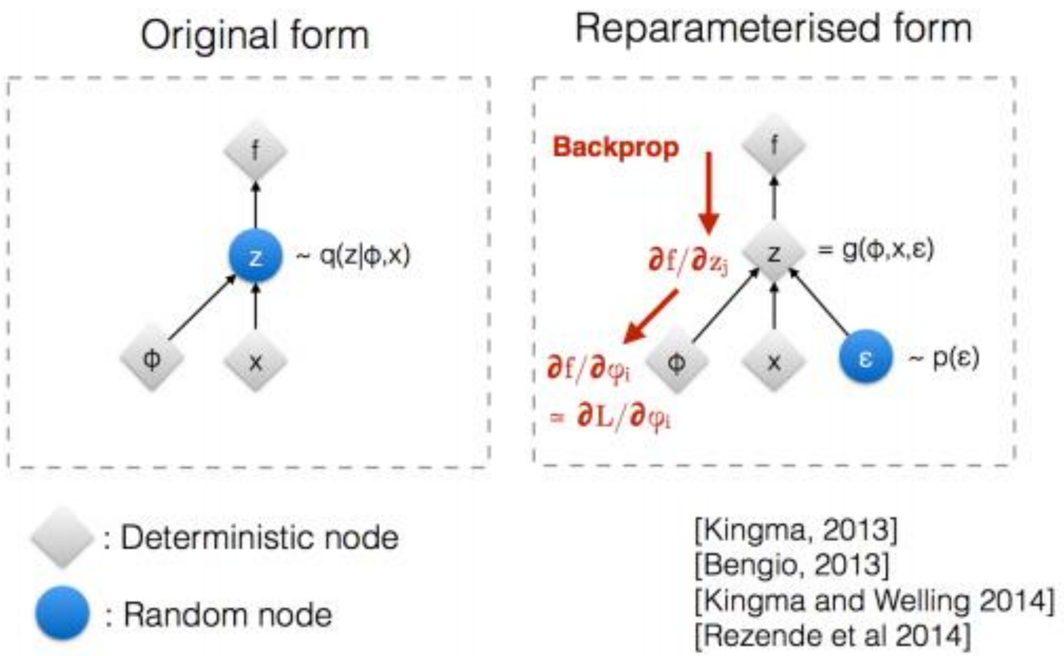


Figure is borrowed from Carl Doersch. Tutorial on Variational Autoencoders

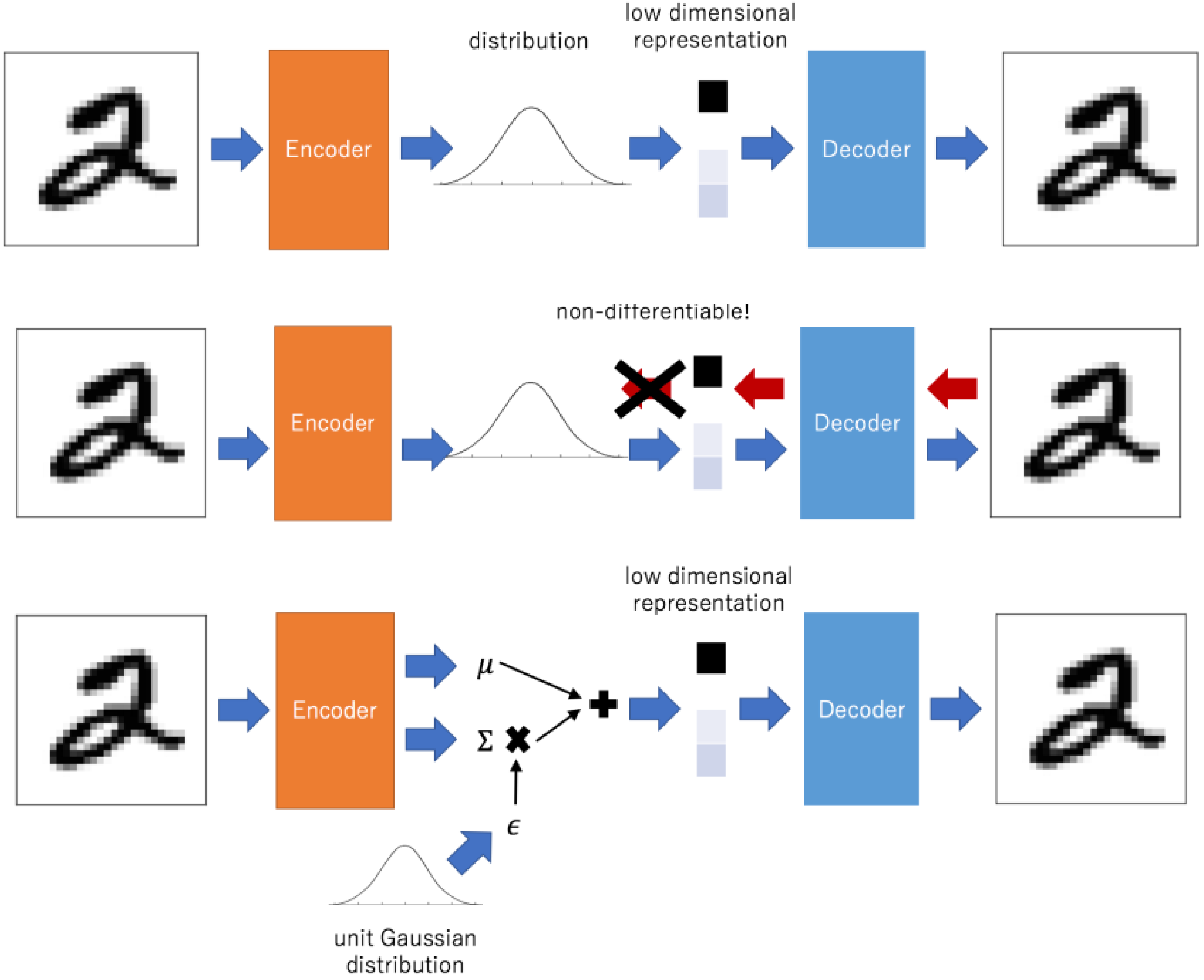


- Left: sampling operations ($\mathbf{z}$ from $\mathcal{N}(\mu(\mathbf{x}), \Sigma(\mathbf{x}))$) are *non-differentiable*
- Right: reparameterization trick allows us to be able to do BP operations

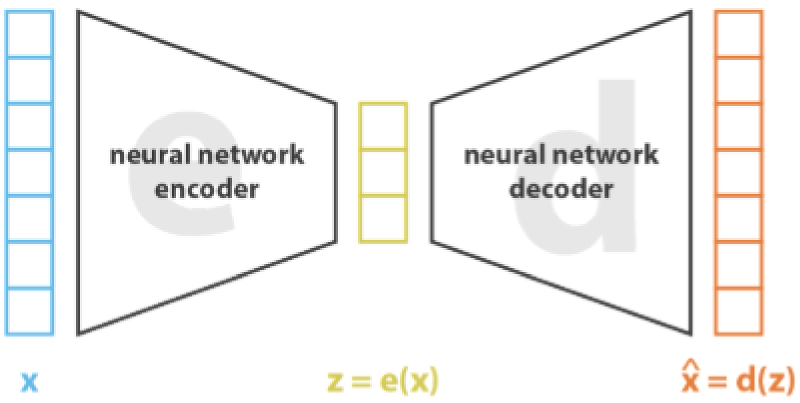
$$z = \mu(\mathbf{x}) + \Sigma^{1/2}(\mathbf{x})\ \varepsilon$$

- $\varepsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{1})$

Variational Auto Encoder

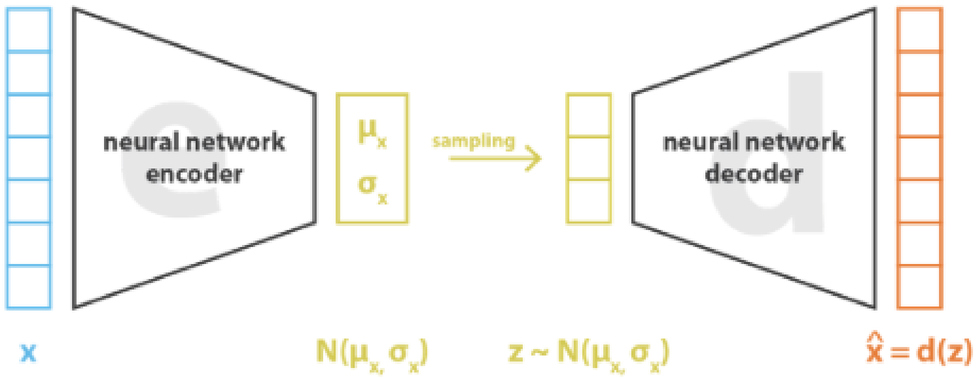


Summary



$$\text{loss} = \| x - \hat{x} \|^2 = \| x - d(z) \|^2 = \| x - d(e(x)) \|^2$$

Auto encoder



$$\text{loss} = \| x - \hat{x} \|^2 + KL[\, N(\mu_x, \sigma_x), N(0, I) \,] = \| x - d(z) \|^2 + KL[\, N(\mu_x, \sigma_x), N(0, I) \,]$$

Variational Auto encoder

reference

https://www.youtube.com/watch?v=KYA-GEhObIs
https://www.youtube.com/watch?v=o_peo6U7IRM
https://lilianweng.github.io/lil-log/2018/08/12/from-autoencoder-to-beta-vae.html#autoencoder

$$\int q(z|x) = 1$$

$$p(x) = \frac{p(z,x)}{p(z|x)}$$

$$E[x] = \int x * f(x)$$

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

$$
\begin{aligned}
\log p_\theta(x^{(i)}) &= \mathbf{E}_{z\sim q_\phi(z|x^{(i)})}\left[\log p_\theta(x^{(i)})\right] \quad (p_\theta(x^{(i)}) \text{ Does not depend on } z) \\
&= \mathbf{E}_z\left[\log \frac{p_\theta(x^{(i)} \mid z)p_\theta(z)}{p_\theta(z \mid x^{(i)})}\right] \quad (\text{Bayes' Rule}) \\
&= \mathbf{E}_z\left[\log \frac{p_\theta(x^{(i)} \mid z)p_\theta(z)}{p_\theta(z \mid x^{(i)})} \frac{q_\phi(z \mid x^{(i)})}{q_\phi(z \mid x^{(i)})}\right] \quad (\text{Multiply by constant}) \\
&= \mathbf{E}_z\left[\log p_\theta(x^{(i)} \mid z)\right] - \mathbf{E}_z\left[\log \frac{q_\phi(z \mid x^{(i)})}{p_\theta(z)}\right] + \mathbf{E}_z\left[\log \frac{q_\phi(z \mid x^{(i)})}{p_\theta(z \mid x^{(i)})}\right] \quad (\text{Logarithms}) \\
&= \boxed{\underbrace{\mathbf{E}_z\left[\log p_\theta(x^{(i)} \mid z)\right] - D_{KL}(q_\phi(z \mid x^{(i)}) \,||\, p_\theta(z))}_{\mathcal{L}(x^{(i)},\theta,\phi)}} + \underbrace{D_{KL}(q_\phi(z \mid x^{(i)}) \,||\, p_\theta(z \mid x^{(i)}))}_{\boxed{\geq 0}}
\end{aligned}
$$

Tractable lower bound which we can take gradient of and optimize! ($p_\theta(x|z)$ differentiable, KL term differentiable)