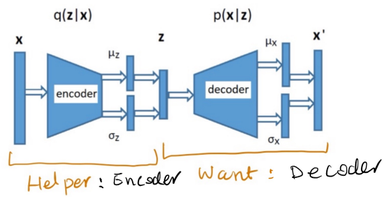
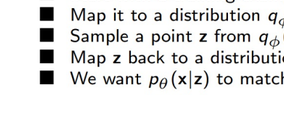
Varitional AutoEncoder

$\Rightarrow \{x_i\}_{i=1}^N$, $p(z) \longrightarrow p(x|z)$



Helper : Encoder Want : Decoder

Training Target : $x' \approx x$



$$L(x_i, \theta, \phi) = E_{z\sim q(z|x_i)}[\log P_\theta(x_i|z)] - D_{KL}(q_\theta(z|x_i)\,\|\,p(z))$$

$$\arg\max_{\theta,\phi} L(\theta,\phi) = \frac{1}{N}\sum_{i=1}^N L(x_i,\theta,\phi)$$

Assume :

$x \longrightarrow z$

(image) (latent vector)

$D_z \ll D_x$

$\begin{cases} z \sim p(z) = \mathcal{N}(0, I) \text{ (Assumption)}, \\ x \sim P_\theta(x|z), \text{ where } \theta: \text{model parameter} \end{cases}$

$P_\theta(x|z) = \mathcal{N}(x|\mu_x(z,\theta), \sigma^2(z,\theta) I)$

Encoder : recognition model $q_\phi(z|x)$

$q_\phi(z|x) = \mathcal{N}(z|\mu_z(x,\phi), \sigma_z^2(x,\phi) I)$

Decoder :

$P_\theta(x|z) = \mathcal{N}(x|\mu_x(z,\theta), \sigma^2(z,\theta) I)$

Objective Function :

$$L(x_i, \theta, \phi) = E_{z\sim q(z|x_i)}[\log P_\theta(x_i|z)] - D_{KL}(q_\theta(z|x_i)\,\|\,p(z))$$

$$\arg\max_{\theta,\phi} L(\theta,\phi) = \frac{1}{N}\sum_{i=1}^N L(x_i,\theta,\phi)$$

- Start with a training example $\mathbf{x}_i$.
- Map it to a distribution $q_\phi(\mathbf{z}|\mathbf{x}_i)$ in the latent space.
- Sample a point $\mathbf{z}$ from $q_\phi(\mathbf{z}|\mathbf{x}_i)$.
- Map $\mathbf{z}$ back to a distribution $p_\theta(\mathbf{x}|\mathbf{z})$ in the data space.
- We want $p_\theta(\mathbf{x}|\mathbf{z})$ to match $\mathbf{x}_i$ as much as possible.

Reparameterization Trick

Why ? $L(x_i, \theta, \phi) = E_{z\sim q(z|x_i)}[\log P_\theta(x_i|z)] - D_{KL}(q_\theta(z|x_i)\,\|\,p(z))$

(reconstruction "error" | Regularization)

Sampling z_{ik} : $z_{i1}, z_{i2} \cdots z_{ik} \sim q_\phi(z|x_i)$

$E_{z\sim q(z|x_i)}[\log P_\theta(x_i|z)] \approx \frac{1}{k}\sum_{i=1}^N \log P_\theta(x_i|z_{ik})$

$\Rightarrow \nabla_\phi L_1 = 0$, where $L_1 = E_{z\sim q(z|x_i)}[\log P_\theta(x_i|z)] \approx \frac{1}{k}\sum_{i=1}^N \log P_\theta(x_i|z_{ik})$

$\Rightarrow$ Cannot process BP !

Reparameterization

$z = \mu_z(x,\phi) + \sigma_z(x,\phi) \odot \varepsilon$, $\varepsilon \in \mathcal{N}(0, I)$

$L_1 = E_{z\sim q_\phi(z|x_i)}[\log P_\theta(x_i|z)]$

$= E_{\varepsilon\sim p(\varepsilon)}[\log P_\theta(x_i|z(x_i,\phi,\varepsilon))] \approx \frac{1}{k}\sum_{i=1}^k \log P_\theta(x_i|z_{ik})$

Where $z_{ik} = \mu_z(x_i,\phi) + \sigma_z(x_i,\phi) \odot \varepsilon_k$, $\varepsilon_k \sim \mathcal{N}(0, I)$

The Second term

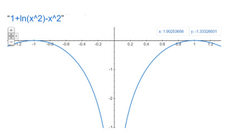
$D_{KL}[q_\phi(z|x_i)\,\|\,p_\theta(z)]$

$= \frac{1}{2}\sum_{j=1}^J (1 + \log(\sigma_{ij}^2) - \mu_{ij}^2 - \sigma_{ij}^2)$

where : J : dimension of z , σ_{ij} : j-th Component of $\sigma_z(x_i,\phi)$

μ_{ij} : j-th Component of $\mu_z(x_i,\phi)$

- The gradient $\nabla_\phi L_2$ is straightforward compute



Final Fuction :

$$L(\mathbf{x}_i,\theta,\phi) \approx \frac{1}{K}\sum_{k=1}^K \log p_\theta(\mathbf{x}_i|\mathbf{z}_{ik}) + \frac{1}{2}\sum_{j=1}^J \left(1 + \log(\sigma_{ij}^2) - \mu_{ij}^2 - \sigma_{ij}^2\right)$$

where $\mathbf{z}_{ik} = \mu_z(\mathbf{x}_i,\phi) + \sigma_z^2(\mathbf{x}_i,\phi) \odot \epsilon_k$, and $\epsilon_k \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.

- We have discussed how to compute the gradient $\nabla_{\theta,\phi} L$. Using the gradient, we can estimate θ and ϕ simultaneously using gradient ascent.
- The hyperparameter K is usually set to 1.

The Theory of VAE

$P(x) \xrightarrow{\text{Sampling}} \{x_i\}_{i=1}^N \xrightarrow{\text{learning}} Q(x)$

Obj: $\min D_{KL}(P\|Q)$

$\Leftrightarrow \min H(P,Q) = -\int P(x)\log Q(x)\,dx = -\frac{1}{N}\sum_{i=1}^N \log Q(x_i)$ (cross entropy)

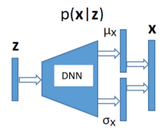
$\Leftrightarrow \max L(P,Q) = \int P(x)\log Q(x)\,dx = \frac{1}{N}\sum_{i=1}^N \log Q(x_i)$ (log likelihood)

Hard to maximize

VAE maximize a Lower bound of the Likelihood

Model : $z_i \sim p(z)$ $x_i \sim p_\theta(x|z_i)$

Why ?



$p_\theta(\mathbf{x}|\mathbf{z}) = \mathcal{N}(\mathbf{x}|\mu_x(\mathbf{z},\theta), \sigma_x^2(\mathbf{z},\theta)\mathbf{I})$

$\max_\theta \frac{1}{N}\sum_{i=1}^N \log P_\theta(x_i)$, $P_\theta(x_i) = \int P(x_i|z)P(z)\,dz$ 生成模型公式

$\approx \frac{1}{K}\sum_{i=1}^K P_\theta(x_i|z_k)$

$= \prod_{j=1}^D e^{-\frac{(x_{ij}-\mu_{kj})^2}{2\sigma_{kj}^2}}$ $\quad x_i = (x_{i1}, \cdots, x_{iD})$

for image : $D = 256\times256$ or larger

$= 0.99 \times 0.98 \times \cdots \times 0.99 = 0$

$$\begin{aligned} 0 \le \mathcal{D}_{KL}(q(\mathbf{z}|\mathbf{x})||p(\mathbf{z}|\mathbf{x})) &= \int q(\mathbf{z}|\mathbf{x})\log\frac{q(\mathbf{z}|\mathbf{x})}{p(\mathbf{z}|\mathbf{x})}d\mathbf{z} \\ &= \int q(\mathbf{z}|\mathbf{x})\log\frac{q(\mathbf{z}|\mathbf{x})p(\mathbf{x})}{p(\mathbf{z}|\mathbf{x})p(\mathbf{x})}d\mathbf{z} \\ &= \int q(\mathbf{z}|\mathbf{x})\log\frac{q(\mathbf{z}|\mathbf{x})p(\mathbf{x})}{p(\mathbf{z},\mathbf{x})}d\mathbf{z} = \int q(\mathbf{z}|\mathbf{x})\log\frac{q(\mathbf{z}|\mathbf{x})p(\mathbf{x})}{p(\mathbf{x}|\mathbf{z})p(\mathbf{z})}d\mathbf{z} \\ &= \int q(\mathbf{z}|\mathbf{x})\log p(\mathbf{x})d\mathbf{z} + \int q(\mathbf{z}|\mathbf{x})\log\frac{1}{p(\mathbf{x}|\mathbf{z})}d\mathbf{z} + \int q(\mathbf{z}|\mathbf{x})\log\frac{q(\mathbf{z}|\mathbf{x})}{p(\mathbf{z})}d\mathbf{z} \\ &= \log p(\mathbf{x}) - E_{z\sim q(z|x)}[\log p(z|x)] + \mathcal{D}_{KL}(q(\mathbf{z}|\mathbf{x})||p(\mathbf{z})) \end{aligned}$$

Therefore,

$$E_{z\sim q(z|x)}[\log p(z|x)] - \mathcal{D}_{KL}(q(\mathbf{z}|\mathbf{x})||p(\mathbf{z})) \le \log p(\mathbf{x})$$

The LHS is called the **variational lower bound (VLB)** or the **evidence lower bound (ELBO)** of the per-sample loglilihood $\log p(\mathbf{x})$ Equality holds when $q(\mathbf{z}|\mathbf{x}) = p(\mathbf{z}|\mathbf{x})$.

$$\log p_\theta(\mathbf{x}_i) \approx \log\frac{1}{K}\sum_{k=1}^K p_\theta(\mathbf{x}_i|\mathbf{z}_k) \tag{1}$$

where the values of $\mathbf{z}$ are sampled from the prior $p(\mathbf{z})$. Those values do not depend on $\mathbf{x}_i$ and *negligible probabilities to* $\mathbf{x}_i$. The RHS is a poor approximation of $\log p_\theta(\mathbf{x}_i)$.

- Here is the VAE objective function is a lower bound of $\log p_\theta(\mathbf{x}_i)$:

$$\frac{1}{K}\sum_{k=1}^K \log p_\theta(\mathbf{x}_i|\mathbf{z}_{ik}) + \frac{1}{2}\sum_{j=1}^J \left(1 + \log(\sigma_{ij}^2) - \mu_{ij}^2 - \sigma_{ij}^2\right) \le \log p_\theta(\mathbf{x}_i) \tag{2}$$

where the values of $\mathbf{z}_{ik}$ are sampled in such a way that they depend upon $\mathbf{x}_i$. *Those values usually give high probabilities to* $\mathbf{x}_i$. The LHS is a better approximation (lower bound) of $\log p_\theta(\mathbf{x}_i)$.