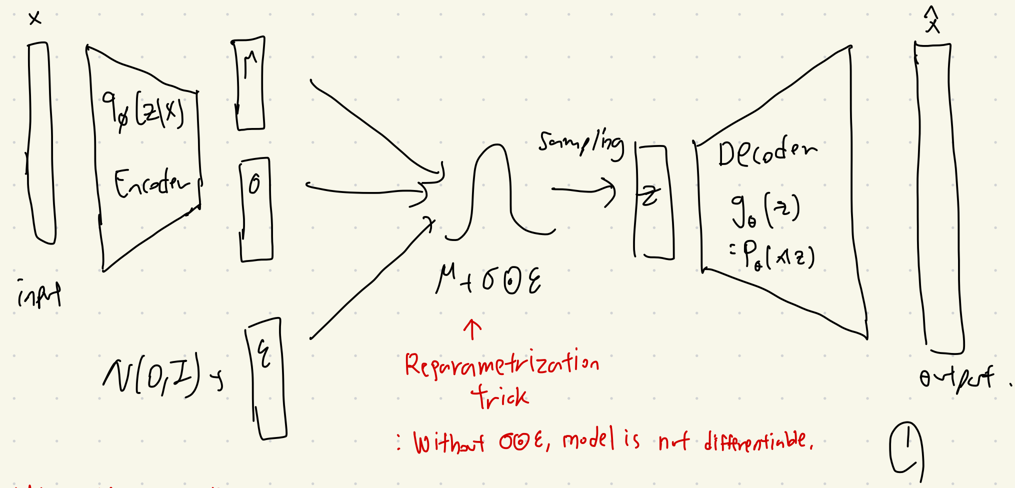
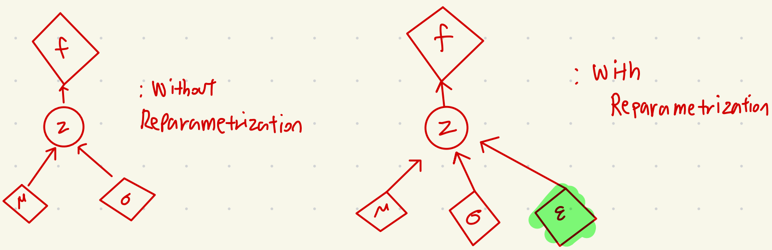
VAE structure:



Visualizing Reparametrization



Objective of VAE: Learn model parameters to maximize likelihood of training data : $P_\theta(X)$

And, $P_\theta(X) = \int P_\theta(z) P_\theta(x|z) dz$

Why? $\frac{P_\theta(X,z)}{P_\theta(z)} = P_\theta(X|z) \Rightarrow P_\theta(X,z) = P_\theta(z) P_\theta(x|z)$

And $\int P_\theta(X,z) dz = P_\theta(X)$.

Hence, $P_\theta(x) = \int P_\theta(z) P_\theta(x|z) dz$

However, $\int P_\theta(x|z) dz$ is intractable

It implies we would derive a lower bound on the data likelihood, which is tractable

Let's work on log data likelihood

$$\log P_\theta(x^{(i)}) = \mathbb{E}_{z \sim q_\phi(z|x)} \left[\log P_\theta(x^{(i)}) \right]$$

$$= \mathbb{E}_{z \sim q_\phi(z|x)} \left[\log \frac{P_\theta(x^{(i)}|z) P_\theta(z)}{P_\theta(z|x^{(i)})} \right]$$

★ Use Bayes' Rule

$P(z|x) = \frac{P(x|z)P(z)}{P(x)} \Rightarrow P(x) = \frac{P(x|z)P(z)}{P(z|x)}$

$$= \mathbb{E}_{z\sim q_\phi(z|x)}\left[\log\left(\frac{p_\theta(x^{(i)}|z)\,p_\theta(z)}{p_\theta(z|x^{(i)})}\cdot\frac{q_\phi(z|x^{(i)})}{q_\phi(z|x^{(i)})}\right)\right]$$

= Multiplying 1

$$= \mathbb{E}_{z\sim q_\phi(z|x)}\left[\log\left(p_\theta(x^{(i)}|z)\cdot\frac{p_\theta(z)}{q_\phi(z|x^{(i)})}\cdot\frac{q_\phi(z|x^{(i)})}{p_\theta(z|x^{(i)})}\right)\right]$$

$$= \mathbb{E}_{z\sim q_\phi(z|x)}\left[\log\left(p_\theta(x^{(i)}|z)\right)-\log\left(\frac{q_\phi(z|x^{(i)})}{p_\theta(z)}\right)+\log\left(\frac{q_\phi(z|x^{(i)})}{p_\theta(z|x^{(i)})}\right)\right]$$

$$= \mathbb{E}_z\left[\log p_\theta(x^{(i)}|z)\right]-\mathbb{E}_z\left[\log\left(\frac{q_\phi(z|x^{(i)})}{p_\theta(z)}\right)\right]+\mathbb{E}_z\left[\log\left(\frac{q_\phi(z|x^{(i)})}{p_\theta(z|x^{(i)})}\right)\right]$$

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

$$\star\ \mathbb{E}_{z\sim q_\phi(z|x)}\left[\log\frac{q_\phi(z|x^{(i)})}{p_\theta(z)}\right]=\int q_\phi(z|x^{(i)})\log\left(\frac{q_\phi(z|x^{(i)})}{p_\theta(z)}\right)dz = KL(P\|Q)$$

Since $KL(P\|Q)=\sum_x P(x)\log\frac{P(x)}{Q(x)}$

However, we know $p_\theta(z|x^{(i)})$ is intractable. However KL always ≥ 0.
Hence,

$$\underbrace{\mathbb{E}_{z\sim q_\phi(z|x)}\left[\log p_\theta(x^{(i)}|z)\right]-KL\left(q_\phi(z|x^{(i)})\,\|\,p_\theta(z)\right)}_{\mathcal{L}(x^{(i)},\theta,\phi)}$$ is our ELBO.

$\mathcal{L}(x^{(i)},\theta,\phi)$: Tractable lower bound.

Hence, VAE Objective : $\theta^*,\phi^* = \underset{\theta,\phi}{\arg\max}\sum_{i=1}^{N}\mathcal{L}(x^{(i)},\theta,\phi)$

Reconstruction Error Regularization

$$\theta^*,\phi^* = \underset{\theta,\phi}{\arg\min}\sum_{i=1}^{N}-\mathcal{L}(x^{(i)},\theta,\phi)=\sum_{i=1}^{N}-\log p_\theta(x^{(i)}|g_\theta(z))+KL\left(q_\phi(z|x^{(i)})\,\|\,p(z)\right)$$

where $p(z)=N(0,I)$

$\Downarrow$

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

Regularization:

Assumption 1: $q_\phi(z|x^{(i)})\sim N(\mu^{(i)},\sigma^{(i)}I)$

Assumption 2: $p_\theta(z)\sim N(0,I)$

Cause it makes VAE simple and reasonable.

$$D_{KL}(N_0 \| N_1) = \frac{1}{2}\left(tr(\Sigma_1^{-1}\Sigma_0) + (\mu_1-\mu_0)^T\Sigma_1^{-1}(\mu_1-\mu_0) - k + \ln\left(\frac{\det\Sigma_1}{\det\Sigma_0}\right)\right)$$

Since our $N_1 = N(0, I)$, $\Sigma_1 = I$, $\mu_1 = 0$

Then, $KL(q_\phi(z|x^{(i)}) \| P_\theta(z)) = \frac{1}{2}\sum_{j=1}^{J} \mu_{i,j}^2 + \sigma_{i,j}^2 - \ln(\sigma_{i,j}^2) - 1)$

Reconstruction Error : MSE or Cross Entropy

MSE : When Decoder uses Gaussian Distribution

Cross-Entropy: When Decoder uses Bernoulli Distribution

Therefore, Goal of VAE is

$$\theta^*, \phi^* = \underset{\theta,\phi}{\operatorname{argmin}} \sum_i \underbrace{-\mathbb{E}_{z \sim q_\phi(z|x^{(i)})}\left[\log(P_\theta(x^{(i)}|z))\right]}_{\text{Reconstruction Error (MSE or Cross-Entropy)}} + \underbrace{KL(q_\phi(z|x^{(i)}) \| P(z))}_{\text{Regularization}}$$

: Encouraging a structured and well-behaved latent space

β-VAE (Disentangled VAE)

Recall VAE Loss: $L_{VAE} = \underbrace{-\log P_\theta(x|z)}_{\text{improving decoder } z \to x} + \underbrace{D_{KL}(q_\phi(z|x) \| P_\theta(z))}_{\text{improving encoder (better representation of } z)}$

Another way of writing VAE objective :

$$\max_{\phi,\theta} \mathbb{E}_{x \sim D}\left[\mathbb{E}_{z \sim q_\phi(z|x)} \log P_\theta(x|z)\right]$$

$$\text{s.t } D_{KL}(q_\phi(z|x) \| P_\theta(z)) < \delta$$

: Maximizing probability of generating real data, while keeping distance between real ($P_\theta(z)$) and approximate posterior distribution $(q_\phi(z|x))$ small. (under small constant δ)

- VAE maximization objective can then be rewritten as a Lagrangian with a Lagrangian multiplier β under KKT condition.

$$: \mathbb{E}_{z \sim q_\phi(z|x)} \log P_\theta(x|z) - \beta(D_{KL}(q_\phi(z|x) \| P_\theta(z)) - \delta)$$

$$= \mathbb{E}_{z \sim q_\phi(z|x)} \log P_\theta(x|z) - \beta(D_{KL}(q_\phi(z|x) \| P_\theta(z)) + \beta\delta$$

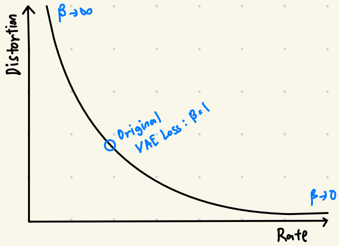
β-VAE Loss hence given by:

$$\mathcal{L}_{Beta}(\phi, \beta) = -\mathbb{E}_{z \sim q_\phi(z|x)} \log p_\theta(x|z) + \beta D_{KL}(q_\phi(z|x) \| p_\theta(z))$$

- When $\beta = 1 \Rightarrow$ Standard VAE
- When $\beta > 1 \Rightarrow$ Stronger constraint on latent bottleneck, follow generative process and thus encourage disentanglement
  ↳ Increasing representation capacity of z. Generally, "stronger disentanglement" means more Gaussian like for most linear definition.
- Disadvantage: Trade-off between disentanglement and reconstruction capability.

★ Distortion: $d(x, \hat{x})$ where $\hat{x}$ reconstruction/estimate of x given z. It's a reconstruction loss. e.g. $-\log(p(x|z)) = MSE[x|z]$

★ Rate: Regularization: $KL(q_\phi(z|x) \| p(z))$



As $\beta \to 0$, Optimizer Prioritizes minimizing distortion: Rate ↑, Distortion ↓

$\beta \to \infty$, Optimizer prioritizes minimizing rate over maximizing distortion

Rate ↓, Distortion ↑