No.

Date

Generative Model:

Sample new points from a model which learns the true distribution based on random samples from that true distribution

Variational Autoencoder design that $p_\theta(x,z)$ where latent variable $z \sim N(0,I)$

why Gaussian?
1. consistency among transformation
2. its the limit of many distributions

covariance matrix is I

Evidence Lower Bound (ELBO, Variational Lower Bound)

Objective: Get $\log p_\theta(x)$

$$\log p_\theta(x) = \log \int_z p_\theta(x,z)\,dz$$

$\mu(z)$ is an arbitrary distribution

$$= \log \int_z p_\theta(x,z)\frac{\mu(z)}{\mu(z)}\,dz$$

$$= \log E_{z\sim\mu(\cdot)} \frac{p_\theta(x,z)}{\mu(z)}$$

According to Jensen's Inequality, since $\log(x)$ is concave,

$$\geq E_{z\sim\mu(\cdot)} \log \frac{p_\theta(x,z)}{\mu(z)}$$

$p_\theta(z) = \prod_{i=1}^{n} p_\theta(z_{i-1}|z_i)$

$$= E_{z\sim\mu(\cdot)} \log \frac{p_\theta(x|z)\,p_\theta(z)}{\mu(z)}$$

$\mu(z) = \hat{\prod}_{i=1} \mu(z_i|z_{i-1})$

$$= E_{z\sim\mu(\cdot)}\left[\log p_\theta(x|z) + \log \frac{p_\theta(z)}{\mu(z)}\right]$$

$$= E_{z\sim\mu(\cdot)} \log p_\theta(x|z) - E_{z\sim\mu(\cdot)} \log \frac{\mu(z)}{p_\theta(z)}$$

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

Reconstruction Loss when sampling $N(\text{Decoder}_\theta(z), I)$ from $\rightarrow \|\hat{x}-x\|_2^2$ MSE

$D_{KL}(\mu(\cdot)\,\|\,p_\theta(\cdot))$ Kind of a regularization

No.

Date

To understand how it learns both mapping of $x \to z$ and $z \to x$

The lower bound $E_{z\sim\mu(\cdot)} \log \frac{P_\theta(z|x)P_\theta(x)}{\mu(z)}$

$$= E_{z\sim\mu(\cdot)} \log P_\theta(x) + E_{z\sim\mu(\cdot)} \log \frac{P_\theta(z|x)}{\mu(z)}$$

not related to θ, x, so basically identity here

$$= E_{z\sim\mu(\cdot)} \log P_\theta(x) - D_{KL}(\mu(z) \| P_\theta(z|x))$$

original object — hard to directly compute, so we use $\mu(\cdot)$

For VAE $\mu(\cdot) = q_\phi(z|x)$

$$= \begin{cases} \theta e^{-\theta z}, & z \geq 0 \\ 0, & z < 0 \end{cases} = \theta e^{-\theta z} \mathbb{I}(z \geq 0)$$

$\sim N(\hat{\mu}(x), \hat{\sigma}^2(x))$ or $N(x; \hat{\mu}, \hat{\sigma}^2 I)$

what your network computes

$z = \hat{\mu}(x) + \hat{\sigma}^2(x)\,\varepsilon, \quad \varepsilon \sim N(0, I)$

Reparameterization transfer sampling into non-random z so that we can use gradient

* D_{KL} between two Gaussian Distributions $D_{KL}(u\|v)$ $\quad u: \mu_1, \sigma_1^2$ $\quad v: \mu_2, \sigma_2^2$

$$D_{KL}(u\|v) = E_{x\sim u(\cdot)} \log \frac{u(x)}{v(x)}$$

$$= E_{x\sim u(\cdot)} \log u(x) - \log v(x)$$

$$= E_{x\sim u(\cdot)} \log \left[\frac{1}{\sqrt{2\pi}\sigma_1} e^{\frac{(x-\mu_1)^2}{-2\sigma_1^2}}\right] - \log \left[\frac{1}{\sqrt{2\pi}\sigma_2} e^{\frac{(x-\mu_2)^2}{-2\sigma_2^2}}\right]$$

$$= E_{x\sim u(\cdot)} \log \left(\frac{\sigma_2}{\sigma_1}\right) + \frac{1}{2\sigma_2^2}(x-\mu_2)^2 + \frac{1}{2\sigma_1^2}(x-\mu_1)^2$$

$\because E_{x\sim u(\cdot)}(x-\mu_1)^2 = \sigma_1^2$

$$E_{x\sim u(\cdot)}(x-\mu_2)^2 = E_{x\sim u(\cdot)}(x-\mu_1+\mu_1-\mu_2)^2$$

$$= \sigma_1^2 + (\mu_1-\mu_2)^2 + 0$$

$\therefore D_{KL}(u\|v) = \log\left(\frac{\sigma_2}{\sigma_1}\right) + \frac{\sigma_1^2 + (\mu_1-\mu_2)^2}{2\sigma_2^2} - \frac{1}{2}$ (closed-form solution)

No.

Date

Likelihood function: $L(\theta|x) = P(X=x|\theta)$

$D_{KL}(P||Q) = \int_{-\infty}^{\infty} p(x) \log\left(\frac{p(x)}{q(x)}\right) dx$

① $D_{KL}(P||Q) \neq D_{KL}(Q||P)$

② $D_{KL}(P||Q) \geq 0$, $=0$ only when $P=Q$

Diffusion Model – DDPM

with latent variable $z_1, z_2 \dots z_n$,

we want to generate data with Markov chain

$p_\theta(x, z_1, \dots z_n) = p_\theta(z_n)(\prod_{i=2:n} p_\theta(z_{i-1}|z_i))\, p_\theta(x|z_i)$

similar to sample n times with VAE

where $p_\theta(z_{i-1}|z_i) = N(z_{i-1}; \mu_\theta(z_i,t), \Sigma_\theta(z_i,t))$

Human predefined mapping such configuration makes scaling stable

forward → $q(z_i|x, z_1, \dots z_{i-1}) = q(z_i|z_{i-1}) = N(z_i, \frac{a_i}{a_{i-1}} z_{i-1}, (1-(\frac{a_i}{a_{i-1}})^2) I)$

$\sqrt{1-\beta_i}$ ☆ β_i

backward →

Since $q(z_{i-1}|z_i, x) = \frac{q(z_{i-1}, z_i, x)}{q(z_i, x)} = \frac{q(z_i|z_{i-1}, x)\, q(z_{i-1}|x)\, q(x)}{q(z_i|x)\, q(x)}$

$\propto \exp\left(-\frac{1}{2}\left(\frac{(z_i - \sqrt{a_i}z_{i-1})^2}{\beta_i} + \frac{(z_{i-1} - \sqrt{\bar{a}_{i-1}}x)^2}{1-\bar{a}_{i-1}} - \frac{(z_i - \sqrt{\bar{a}_i}x)^2}{1-\bar{a}_i}\right)\right)$

$= \exp\left(-\frac{1}{2}\left(\left(\frac{a_i}{\beta_i} + \frac{1}{1-\bar{a}_{i-1}}\right) z_{i-1}^2 - \left(\frac{2\sqrt{a_i}}{\beta_i} z_i + \frac{2\sqrt{\bar{a}_{i-1}}}{1-\bar{a}_{i-1}} x\right) z_{i-1} + C(z_i, x)\right)\right)$

variance of z_{i-1} mean of z_{i-1} unrelated

which can be deduced in a form $\frac{1}{2a}(x+\frac{b}{2a})^2 + C$ where $\mu = -\frac{b}{2a}$, $\sigma^2 = \frac{1}{a}$.

therefore, $\propto \exp\left(-\frac{1}{2}\frac{(z_{i-1} - \tilde{\mu}(z_i, x))^2}{\sigma^2}\right)$ $\tilde{\mu}$ is linear

KOKUYO

$$\bar{\sigma}^2 = \frac{1}{\frac{\alpha_i}{\beta_i} + \frac{1}{1-\bar{\alpha}_{i-1}}} = \frac{1-\bar{\alpha}_{i-1}}{1-\bar{\alpha}_i} \cdot \beta_i$$

$$\bar{\mu}_i(x_i, x) = \frac{\sqrt{\alpha_i}(1-\bar{\alpha}_{i-1})}{1-\bar{\alpha}_i} x_i + \frac{\sqrt{\bar{\alpha}_{i-1}}\beta_i}{1-\bar{\alpha}_i} x = \frac{1}{\sqrt{\alpha_i}}\left(x_i - \frac{1-\alpha_i}{\sqrt{1-\bar{\alpha}_i}}\varepsilon_i\right) \quad (2)$$

∴ with given x, mean μ is a function of hyperparam, x_i, ε_i (linear)

variance σ^2 is a function of hyperparam

Two tricks

① : Reparameterization

To make sampling from Gaussian differentiable.

for $z \sim N(z; \mu_\theta, \sigma_\theta^2 I)$, we let

$$z = \mu_\theta + \sigma_\theta \odot \varepsilon, \quad \varepsilon \sim N(0, I)$$

randomness is transferred from z to ε

② x_i is a function of x and β $(\alpha_i = 1-\beta_i)$

$$x_i = \sqrt{\alpha_i \alpha_{i-1}}\, x_{i-2} + \sqrt{1-\alpha_i\alpha_{i-1}}\, \bar{\varepsilon}_{i-2}$$

$$= \dots$$

$$= \sqrt{\bar{\alpha}_i}\, x_0 + \sqrt{1-\bar{\alpha}_i}\, \varepsilon \quad (N(0, \sigma_1^2 I) + N(0, \sigma_2^2 I) = N(0, (\sigma_1^2+\sigma_2^2) I))$$

therefore $q(x_i | x) = N(x_i; \sqrt{\bar{\alpha}_i}\, x, (1-\bar{\alpha}_i) I)$

$\sqrt{1-\beta_i}$ makes sure variance $\xrightarrow{n\to\infty}$ 1

$$q(x_{i-1} | x_i, x) = N(x_{i-1}, \tilde{\mu}_i(x_i, x), \bar{\sigma}^2 I)$$

From ELBO in VAE, we know

$$\log p_\theta(x) \geq E_{z\sim\mu(\cdot)} \log \frac{p_\theta(x|z_1)\, p_\theta(z_1)}{\mu(z_1)}$$

$\downarrow \mu \to q$

$$\therefore \quad -\log p_\theta(x) \leq E_q\left[-\log p_\theta(z_n) + \sum_{i=1}^{n} \log \frac{q(z_i|z_{i-1})}{p_\theta(z_{i-1}|z)}\right)$$

$$= E_q\left[-\log p_\theta(z_n) + \sum_{i=2}^{n} \log\left(\frac{q(z_i|z_{i-1})}{p_\theta(z_{i-1}|z_i)}\right) + \log \frac{q(z_1|x)}{p_\theta(x|z_1)}\right]$$

add x to make this computable ☆

$$= E_q\left[-\log p_\theta(z_n) + \sum_{i=2}^{n} \log\left(\frac{q(z_{i-1}|z_i,x)\, q(z_i|x)}{q(z_{i-1}|x)\, p_\theta(z_{i-1}|z_i)}\right) + \log \frac{q(z_1|x)}{p_\theta(x|z_1)}\right]$$

$$= E_q\left[-\log p_\theta(z_n) + \sum_{i=2}^{n} \log\left(\frac{q(z_{i-1}|z_i,x)}{p_\theta(z_{i-1}|z_i)}\right) + \log \frac{q(z_n|x)}{q(z_1|x)} + \log \frac{q(z_1|x)}{p_\theta(x|z_1)}\right]$$

cancel out like $\frac{n}{n-1} \cdot \frac{n-1}{n-2} \cdots \frac{2}{1}$

$$= E_q\left[\log \frac{q(z_n|x)}{p_\theta(z_n)} + \sum_{i=2}^{n} \log\left(\frac{q(z_{i-1}|z_i,x)}{p_\theta(z_{i-1}|z_i)}\right) - \log p_\theta(x|z_1)\right]$$

$$= E_q\left[D_{KL}(q(z_n|x)\,\|\,p_\theta(z_n)) + \sum_{i=2}^{n} D_{KL}(q(z_{i-1}|z_i,x)\,\|\,p_\theta(z_{i-1}|z_i)) - \log p_\theta(x|z_1)\right]$$

- $D_{KL}(q(z_n|x)\,\|\,p_\theta(z_n))$: we don't optimize this, treat as constant
- $p_\theta(z_n)$: pure noise, $N(0, I)$
- $p_\theta(z_{i-1}|z_i)$: used for generation; make it learn without data x
- $-\log p_\theta(x|z_1)$: simple reconstruct

Recall if p, q are both Gaussian

$$KL(p,q) = \log \frac{\sigma_2}{\sigma_1} + \frac{\sigma_1^2 + (\mu_1-\mu_2)^2}{2\sigma_2^2} - \frac{1}{2}$$

By design, the model learn from a bunch of latent variables to match q and p

No.

Date

Now that 1. $q(z_{i-1} | z_i, x) = N(z_{i-1}, \hat{\mu}_i(z_i, x), \tilde{\delta}^2 I)$

where $\hat{\mu}_i = \frac{1}{\sqrt{a_i}}(z_i - \frac{1-a_i}{\sqrt{1-\bar{a}_i}}\varepsilon_i)$

$\hat{\delta}^2 = \tilde{\beta}_i = \frac{1-\bar{a}_{i-1}}{1-\bar{a}_i}\beta_i$

2. $p_\theta(z_{i-1} | z_i) = N(z_{i-1}; \mu_\theta(z_i, i), \Sigma_\theta(z_i, i))$

$\therefore$ Loss $= E_q[D_{KL}(q(z_{i-1} | z_i, x) || p(z_{i-1} | z_i))]$

discarded since unrelated

$= E_q[\| \hat{\mu}_i(z_i, x) - \mu_\theta(z_i, i) \|_2^2] \times \frac{1}{2\tilde{\delta}_i^2} + C$

$= \frac{1}{2} E_\varepsilon[(SNR(i-1) - SNR(i)) \| x - \hat{x}_\theta(z_i, i) \|_2^2]$ match x and predicted x

$SNR = \frac{\bar{a}_i}{\tilde{\delta}_i^2 \leftarrow 1-\bar{a}_i}$ in ddpm

noise rate

we then define $\mu_\theta(z_i, i) = \frac{1}{\sqrt{a_i}}(z_i - \frac{1-a_i}{\sqrt{1-\bar{a}_i}}\varepsilon_\theta(z_i, i))$

we predict the noise

Objective: $E_{x,\varepsilon}[\|\varepsilon - \varepsilon_\theta(z_i, i)\|^2]$ $\varepsilon \sim N(0, I)$

What if infinite z_i / continuous time?

$SNR'(t) = \frac{d\,SNR(t)}{dt}$

$\therefore$ Objective becomes: $-\frac{1}{2} E_\varepsilon[\int_0^1 SNR'(t) \| x - \hat{x}_\theta(z_t, t) \|_2^2 dt]$

SDE perspective:

Forward: $dx = f(x, t)dt + g(t)dw$ (drift; weiner motion (random); steps in vector space, describe how distribution is changing)

Reverse: $dx = [f(x, t) - g^2(t)\nabla_x \log p_t(x)]dt + g(t)d\bar{w}$

score of distribution (gradient of log prob)

Using SDE solver gives us more accurate prediction than discrete version

Campus we can view ddpm in a da perspective or score match perspective