DIFFUSION MODELS

- DM is a generative model
  ↳ RECALL (other) Gen. Models
  VAE, GAN, ...

- Problem Formulation (Basic case)

Training Set : $\{ x^{(1)}, \dots, x^{(N)} \} \sim q_{data}(x)$

↳ images with an unknown underlying distribution q

GOAL : learn generative model $p_\theta(x)$

s.t. $p_\theta(x) \sim q_{data}(x)$

i.e. learn a parametric model to approximate the unknown underlying distribution q

+ we want easy sampling from p
  (want to generate new images)

* Compared to VAE / GAN, DM uses a different approach for $x = G(z)$

KEY IDEA : add & remove noise from images
learn p_θ from this add / remove process

/clean
training /original image
(known)

~ q_{data}
(unknown)

as you add more noise
...

totally noisy

$$\boxed{q(x_0)}$$

original noiseless distribution

$q(x_1)$ ... $q(x_T)$

FORWARD PROCESS
"noising"

$q(x_0)$: original distribution

$q(x_T)$: noise distribution (after T steps of "noising")

at each step of FWD process, we add small amount of noise

Q: do we know the noise distribution ?

Q: if we have a good noise model, can we characterize $q(x_1)$ in terms of $q(x_0)$?

$$q(x_t | x_{t-1}) \sim N(\sqrt{1-\beta_t}\, x_{t-1}, \beta_t I)$$

mean

variance

if noise is Gaussian,
mean is func. of previous sample
(scaled by $\sqrt{1-\beta_t}$)
var is β_t times Identity

in other words:

$$\boxed{X_t = \sqrt{1-\beta_t}\, X_{t-1} + \sqrt{\beta_t}\, \mathcal{E} \qquad \mathcal{E} \sim N(0, I)}$$

β_t : coeff

why ?

ensures that variance is preserved (doesn't explode)

Q: do we need to sequentially generate X_t ? or can we write X_T directly ?

↳ it's not deterministic because $\mathcal{E}$ is sampled from $N(0, I)$
(it's not the same everytime)

$$X_1 = \sqrt{1-\beta_1}\, X_0 + \sqrt{\beta_1}\, \mathcal{E}_0$$

$$X_2 = \sqrt{1-\beta_2}\, X_1 + \sqrt{\beta_2}\, \mathcal{E}_1$$

$$= \underbrace{\sqrt{1-\beta_2}}_{\sqrt{\alpha_2}}\, \underbrace{\sqrt{1-\beta_1}}_{\sqrt{\alpha_1}}\, X_0 + \sqrt{1-\beta_2}\sqrt{\beta_1}\, \mathcal{E}_0 + \sqrt{\beta_2}\, \mathcal{E}_1$$

Gaussian $N\left(0, \left[(1-\beta_2)\beta_1 + \beta_2\right] I\right)$

$N(0, (1-\bar{\alpha}_2) I)$

define : $1-\beta_2 = \alpha_2$, $1-\beta_1 = \alpha_1$

combined $\bar{\alpha}_2 = \alpha_1 \alpha_2 = (1-\beta_1)(1-\beta_2)$

$$\bar{\alpha}_t = \alpha_1 \cdots \alpha_t = \prod_t (1-\beta_t)$$

$$\boxed{\therefore X_2 = \sqrt{\bar{\alpha}_2}\, X_0 + \sqrt{1-\bar{\alpha}_2}\, \mathcal{E} \qquad \mathcal{E} \sim N(0, I)}$$

in general $X_t = \sqrt{\bar{\alpha}_t}\, X_0 + \sqrt{1-\bar{\alpha}_t}\, \mathcal{E}$

"FORWARD PROCESS"

$$\boxed{\begin{aligned} x_t &= \sqrt{\bar{\alpha}_t}\, x_0 + \sqrt{1-\bar{\alpha}_t}\, \varepsilon \quad ; \varepsilon \sim N(0, I) \\ \bar{\alpha}_t &= \alpha_1 \alpha_2 \cdots \alpha_t = (1-\beta_1) \cdots (1-\beta_t) \end{aligned}}$$

note : $\beta_t \in (0,1) \Rightarrow \alpha_t \in (0,1)\ \forall t$

example : assume each $\beta = 1/2$

after $T = 1000$, $\sqrt{\alpha_t} = \left(\left(\frac{1}{2}\right)^{1000}\right)^{1/2}$

$$\therefore x_{1000} = \underbrace{\left(\tfrac{1}{2}\right)^{500}}_{\sim 0} x_0 + \underbrace{\sqrt{1-\left(\tfrac{1}{2}\right)^{500}}}_{\sim 1}\, \varepsilon$$

$$x_{1000} \sim \varepsilon$$

(after a lot of steps, $x_{1000} \rightarrow$ noise)

~~$x \rightarrow N(0, I)$ as~~

as $T \rightarrow \infty$

$$q(x_T \mid x_0) \rightarrow N(0, I)$$

or $\boxed{\lim_{T \rightarrow \infty} q(x_T \mid x_0) = N(0, I)}$

REVERSE
~~BACKWARD~~ PROCESS ?

Can we generate q_{t-1} from q_t ?

(Can we do "denoising" ?)

↳ for now, let's assume we can

if we can,
does it help us get $p_\theta(x)$?

↓
YES ! we can start from noise $\mathcal{E} = X_T$
and start iterative "denoising"
to obtain CLEANER $\hat{X}_{T-1}, \hat{X}_{T-2} \dots$
all the way to $\hat{X}_0$

and $\hat{X}_0 \sim q_0$
↓
distribution of original Data

$$q_{1000} \longrightarrow q_{999} \longrightarrow q_{998} \dots \longrightarrow q_1 \longrightarrow q_0$$

GOAL : $q(x_{t-1} \mid x_t)$

Can we simplify conditional distributions?

BAYES!

GAUSSIAN!

$$q(x_{t-1} \mid x_t) = \frac{q(x_t \mid x_{t-1})\, q(x_{t-1})}{q(x_t)}$$

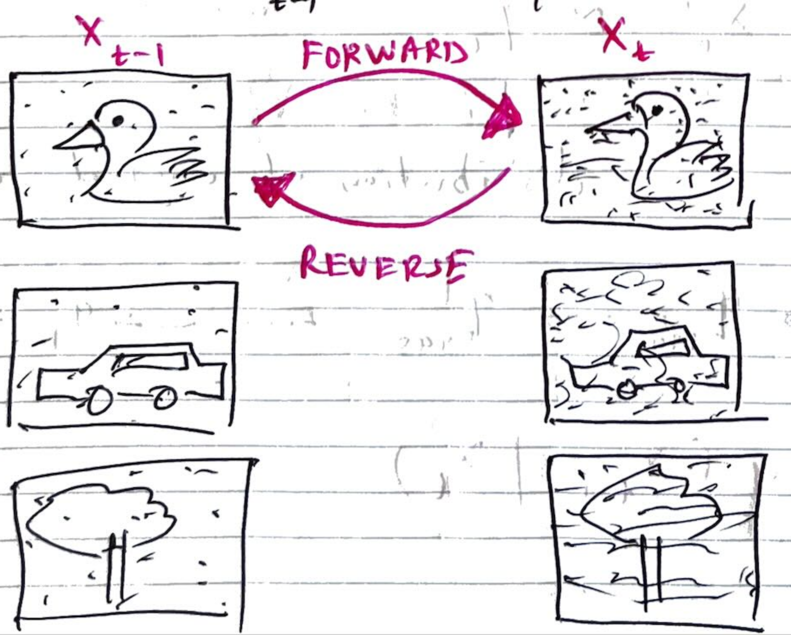
but do we know $q(x_{t-1})$ and $q(x_t)$

↳ not in general, but we know that for high t, $q(x_t) \sim \varepsilon$

but we don't have $q(x_{t-1})$

.... WHAT CAN WE DO?

– consider X_{t-1} that you obtain from FWD process



$q(x_{t-1} \mid x_t) = ?$

"DENOISING"

Recall VAE:

$$z \sim E(X) \qquad \overset{\text{decoder}}{D}(z) = \hat{x}$$

we can model $q(x_{t-1} \mid x_t)$

using a parametric model

$$P_\theta(x_{t-1} \mid x_t)$$

neural network model with weights θ

we will learn θ

P_θ will progressively denoise from X_T = noise

P_θ is the generator

gradually add noise $t-1 \rightarrow t$

find parameters θ that can reverse this.

@ each step

ok. we need to model $P_\theta(x_{t-1} \mid x_t)$

let's model P_θ with a Gaussian

(Gaussians are simple)

$$P_\theta(X_{t-1} \mid X_t) = N(\text{mean}, \text{var})$$

mean : modeled by NN

$$\mu_\theta(X_t, t)$$

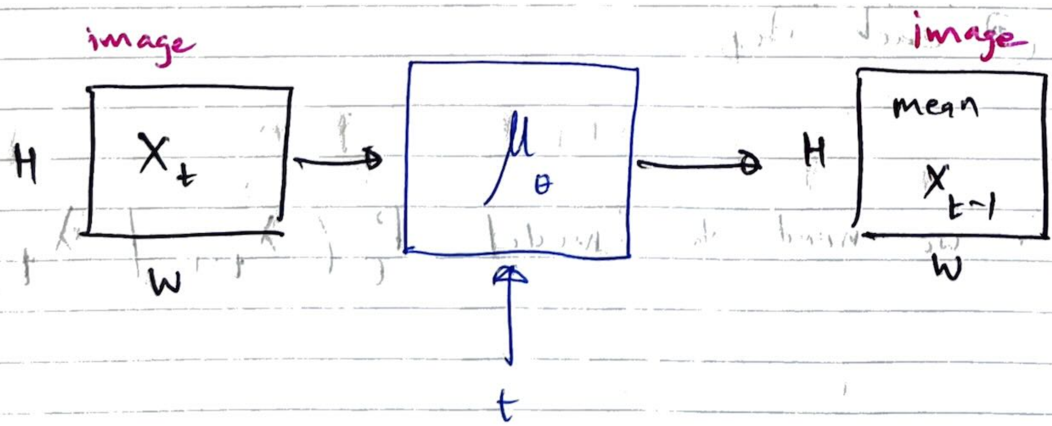
var : some noise

$$\sigma_t^2 I$$

$$P_\theta(X_{t-1} \mid X_t) = N(\mu_\theta(X_t, t), \sigma_t^2 I)$$

does this have to be a function of time?

not necessarily. but often helpful



Now, given this model, we get a generative process

we can express $P_\theta(X_0)$ now $\quad X_0$: clean image

$$\boxed{P_\theta(X_0) = \underbrace{P(X_T)}_{\sim N(0, I)} P_\theta(X_{T-1} | X_T) \cdots P_\theta(X_0 | X_1)}$$

That's it!

SOTA models for image generation, video generation, audio generation ....

One missing piece: how to learn θ?

$$\mu_\theta(x_t, t)$$

just do MLE ...

$$\max_\theta \mathbb{E}_{\substack{x_0 \sim q(x_0) \\ \text{(real data)}}} \left[\log P_\theta(x_0) \right] \longrightarrow \theta^*$$

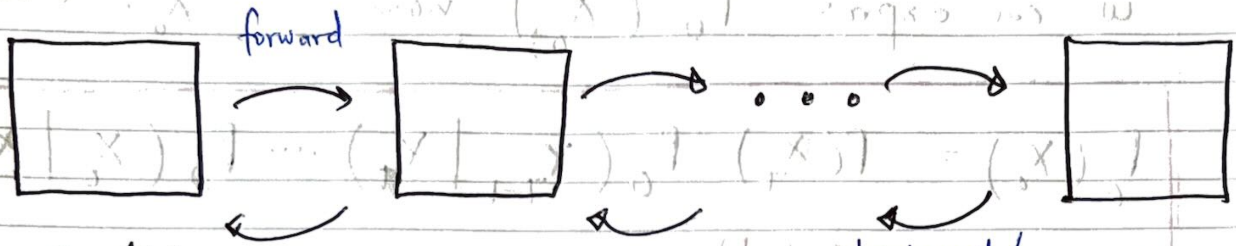
Can we solve this directly? No.

↳ we need to use variational lower bound (ELBO)

Read:
Sohl-Dickstein 2015
Ho et al. 2020

we will not discuss in class

RECAP / SUMMARY of DIFFUSION MODELING



Forward : start with clean image distribution $q(x_0)$
add noise progressively to get $q(x_1) .. q(x_T)$

Reverse : start with $P(x_T) \sim N(0, I)$
model with $\underline{\text{NN } \theta}$ $P_\theta(x_{t-1} \mid x_t) \sim N\left(\underline{\mu_\theta(x_t, t)}, \sigma_t^2 I\right)$

$$\underbrace{\log P_\theta(x) \geq \mathbb{E}_{q(x_{1:T} \mid x_0)} \left[\log \frac{P_\theta(x_{0:T})}{q(x_{1:T} \mid x_0)} \right]}_{\text{ELBO}}$$

$$\theta^* = \min_\theta (- \text{ELBO})$$