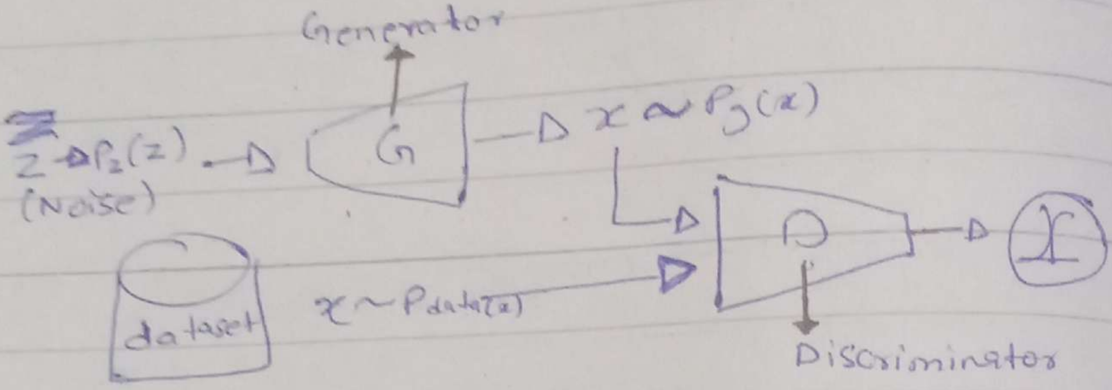
16/04/2025

Generative Adversial Network (GAN)

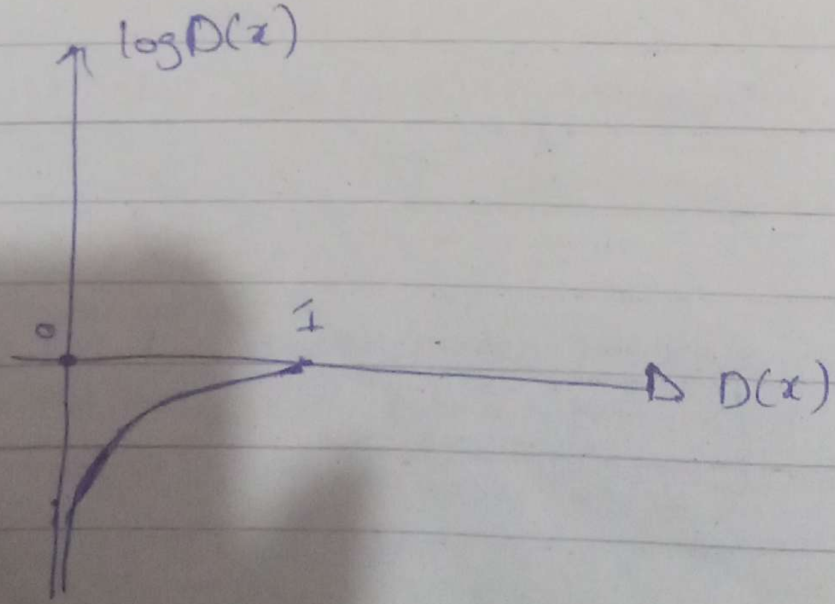
Adversial Learing



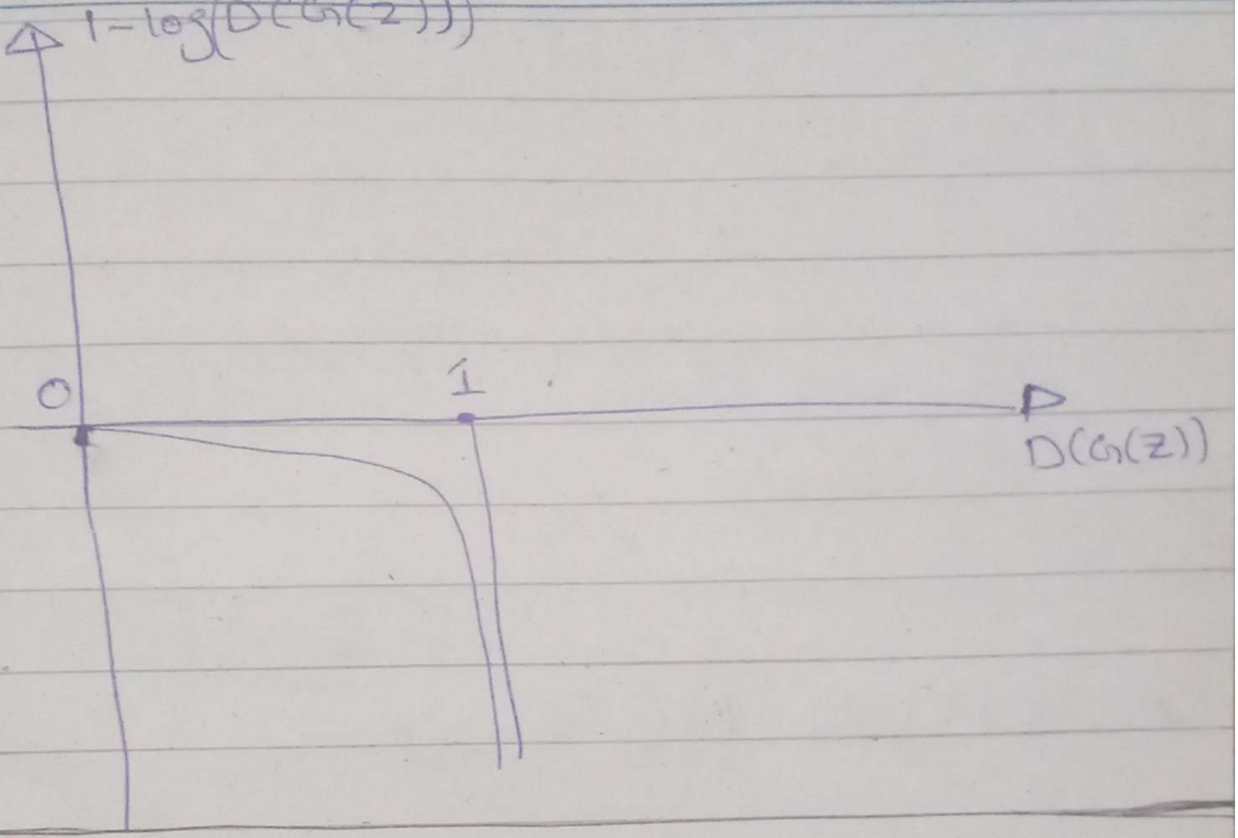
Adversial Learning

$$\min_{G} \max_{D} V(D, G)$$

$$V(D,G) = E_{x \sim P_{data}(x)}[\log D(x)] + E_{z \sim P_z(z)}[\log(1 - D(G(z)))]$$



De convolution $\rightarrow$ ~~Auto~~



Generator's Task is to fool the Discriminator

Training GAN's

$$\max_D V(D,G) \Rightarrow \min_G V(D,G) \Rightarrow \max_D V(D,G)$$
$$\Rightarrow \min_G V(D,G)$$

Diffusion Models:-

Recap of VAE

$$\epsilon \sim N(0,1)$$

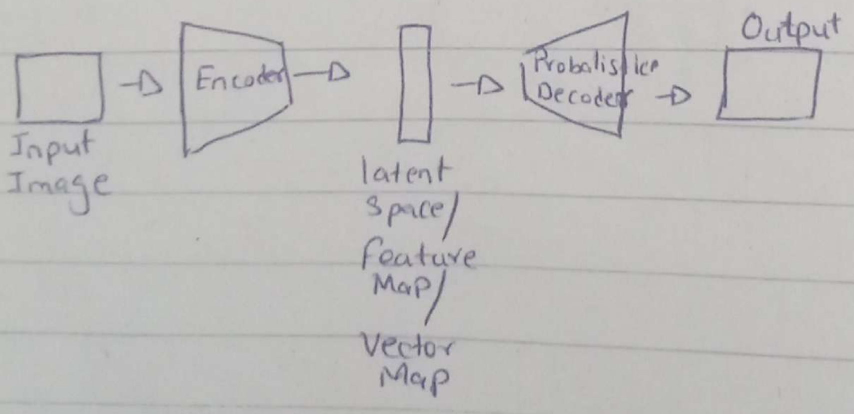
$$Z = E(z) + V(z) \cdot \epsilon$$

We sequentially add noise until we completely distorted the image.

(Noise is always Gaussian)
(Gaussian distribution is the most closet to Natural)

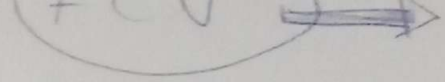
$$x_1 = x_0 + n \quad \text{(Forward)}$$
$$x_0 = x_1 - \hat{n} \quad \text{(Backward)}$$

VAE's is related to Diffussion Model



Variational auto encoder

So Encoder in VAE is similiar to Forward pass in D.M.
& Decoder in VAE is similiar to Backward pass in D.M.



$\boxed{23/04/2025}$

Diffusion Models are trained on noise

$$x^{n} \longrightarrow \boxed{D.M} \longrightarrow \hat{n}$$

n = noise ,

$$x_3 \longrightarrow \boxed{D.M} \longrightarrow \hat{n}$$

$$x_2 = x_3 - \hat{n}$$

Forward Pass

$$q(x_t \mid x_{t-1}) = \mathcal{N}(\underbrace{x_t}_{\text{output}} ; \underbrace{\sqrt{1-\beta_t}\, x_{t-1}}_{\text{Mean}}, \underbrace{\beta_t I}_{\text{Variance (factor of Identity Matrix)}})$$

e.g.

$$\left.\begin{array}{l} x_t = x_{t_1} + x_{t_2} + x_{t_3} \\ \begin{array}{c|ccc} & x_{t_1} & x_{t_2} & x_{t_3} \\ \hline x_{t_1} & \sigma_1 & 0 & 0 \\ x_{t_2} & 0 & \sigma_2 & 0 \\ x_{t_3} & 0 & 0 & \sigma_3 \end{array} \end{array}\right\} \text{Covariance}$$

Forward pass is very simple:
Declare a distributution

Backward pass:

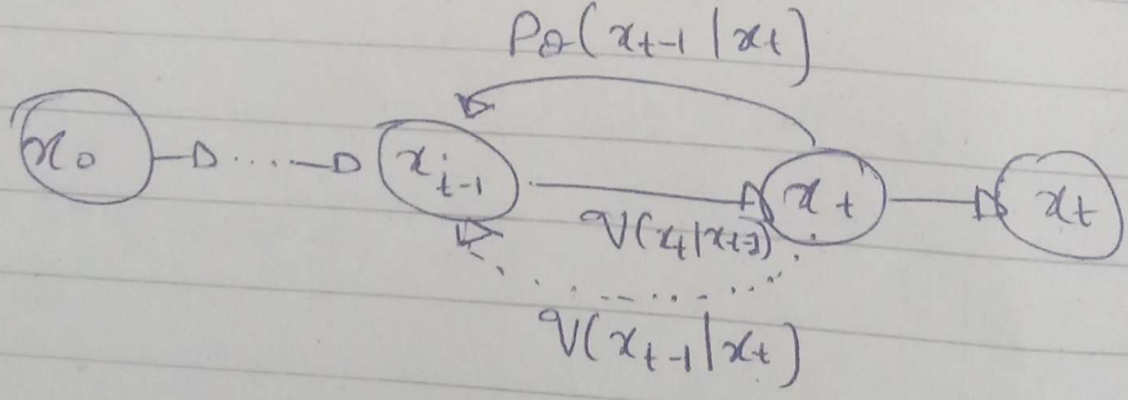
$$P_\theta(x_{t-1}|x_t) = \mathcal{N}(\underbrace{x_{t-1}}_{\text{Outpt}} ; \underbrace{\mu_\theta(x_t,t)}_{\text{Mean}}, \underbrace{\sigma_t^2 I}_{\text{S.D/Variance}})$$

Identity matrix is used to simplify calculations, it reduces relations b/w random variables

Standard Deviation will remain same in FP & BP

but we will try to approximate μ (Mean)

$$\Rightarrow |\mu_n - \mu_{\hat{n}}|$$



07/05/2025

A case study on how to solve the problems

Multimodel Learning

$$\epsilon \sim \mathcal{N}(0,1)$$

$\boxed{24/04/2025}$

$$x_t = \sqrt{\bar{\alpha}_t}\, x_0 + (1-\bar{\alpha}_t)\cdot\epsilon$$

Mean of $q(x_{t-1} | x_t, x_0)$

(Backward Pass)

$$\mu_t = \underbrace{\frac{1}{\sqrt{\alpha_t}}}_{\text{constant}} \left(x_t - \underbrace{\frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}}_{\text{constant}} \underbrace{\epsilon_t}_{\text{Actual Noise}} \right)$$

$\alpha = 1-\beta \rightarrow \begin{bmatrix} \beta_1 = 0.002 \\ \beta_2 = 0.003 \\ \beta_3 = 0.007 \end{bmatrix}$

Δ we endup with

$$\mu_t = x_t - \text{noise}(\epsilon_t)$$

$\downarrow$

This is the noise we added into the image

Actual Mean for the reverse process is

$$\mu_t = x_t - \text{noise}(\epsilon_t)$$

~~If~~ If you have 2 distributions "p" & "q"
And both are ~~are~~ gaussian & you
want to find KL-Divergence there is
a generalized formula for it there.

$$D_{KL}(q \| p) = \frac{1}{2}\left(tr(\underbrace{\Sigma_p^{-1} \Sigma_q}_{\text{Covariance of } p \,\&\, q}) + (\mu_p - \mu_q)^T \Sigma_p^{-1} (\mu_p - \mu_q) - k + \ln\left(\frac{\det \Sigma p}{\det \Sigma q}\right)\right)$$

A

for q =

$$\tilde{\mu}(x_t, x_0) = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{\beta_t}{\sqrt{1-\alpha_t}} \epsilon_t\right)$$

for p =

CamScanner

Work with the transformer layers.

29/04/2025

ATTENTION MECHANISM:-

Drawback

~~Background~~ Very Data hungry (20,000 to 30,000

RNN (Recap)

~~PIEAS~~

The final layers lose the information in the earlier words

feature ~~of~~ vector

Attention Mechanism: It is all about weighted ~~elements~~ average

Context-window

Cross-attention & Self-attention is used in Transformers.

Image net (CNN-based) Vs Transformers (ATTENTION MECHANISM)

30/04/2025

$C_4 = W_1 [\quad] + W_2 [\quad] + W_3 [\quad]$

يحيى شاكر مصر

As it has more importance, it will have more value

weighted Avg

Decoder = I am a ______

Cross-Attention Mechanism

Self-attention → Every word will be checked for context

For self-attention we take concepts from databases. (Query, key, Value)

Every word will be converted to 3

$V_{\text{مصر}} / E_{\text{مصر}} = S_1 V_{\text{مصر}} + S_2 V_{\text{شاكر}} + S_3 V_{\text{يحيى}}$

(Query) مصر | (Key) مصر

(Query) مصر | (Key) شاكر

(Query) مصر | (Key) يحيى

07/05/2025

Paper is called CLIP

→ Contrastive language Image Pre-tr

Learning Transferable Visual Models from Natural language Supervision

Bert used to extract features from Text.

From Features we would used Vision encoder's.

CLIP (Research paper reading is a must)

→ Contrastive Learning Image Pre-training

Text$_1$ → $\boxed{}$ → T_1

$\updownarrow$

Video$_1$ → $\boxed{}$ → I_1

As these 2 belong to the same Sample we put the closer in the feature space.

NSF → Grant

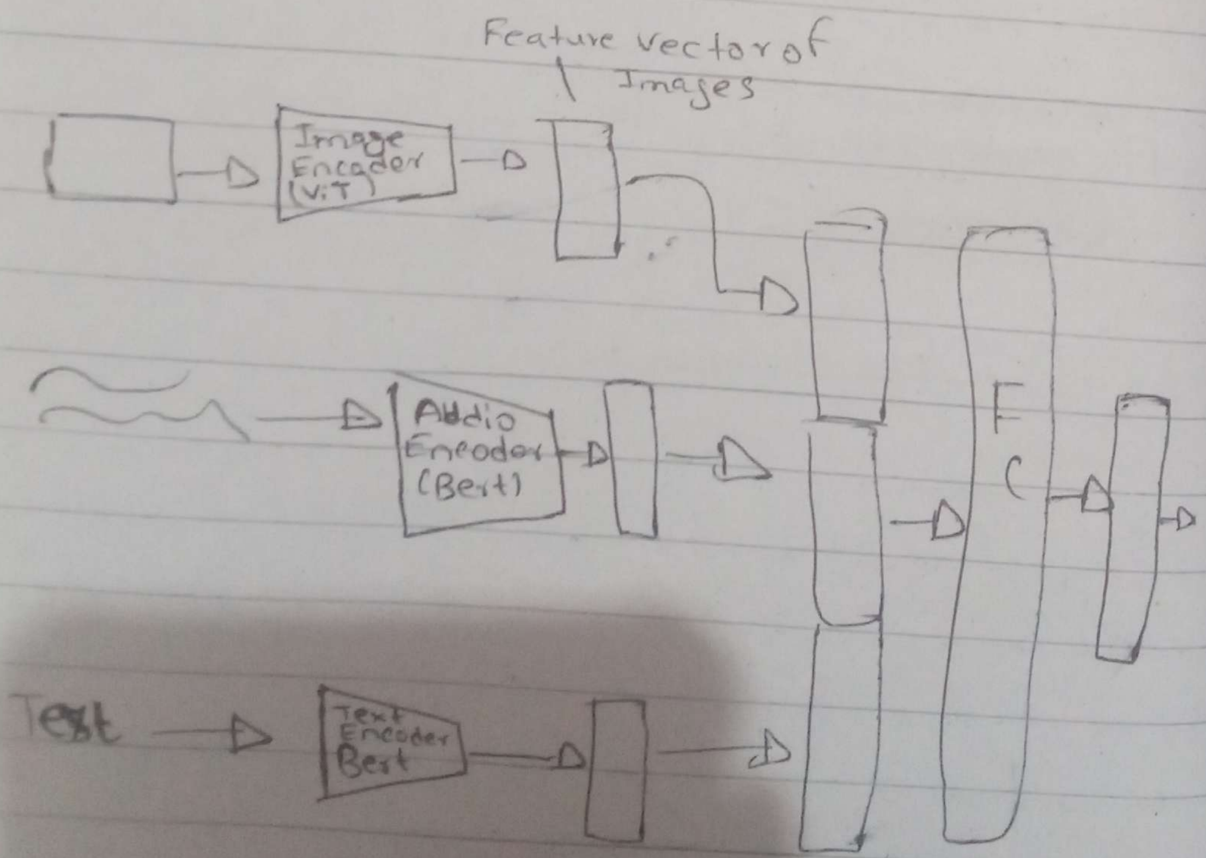
ASR Case Study
→ Automatic Speech Recognition 08/05/2025

ASD

Kids with ~~and~~ autism

Web-based Emotion recognition algo that, takes video as i/p

Text encoder → bert
~~Img~~ Image Encoder → Vit
Audio encoder → bert



08/05/2025

Audio is aligned with the video

Do Frame by frame but it has less accuracy then video based.

3D-CNN for videos but has more parameters ~~but~~ & web-based has limited parameters.

Department of Education
- It's very slow

What is the reason

↳ The Text is not bringing very big impact into the picture, remove it

~~We~~

By applying Fourier Transformation you can ~~app~~ convert 1D signal into 2D signal

As Image encoder is 2D & audio encoder is 1D and ~~we~~ we end up with 2 encoders slowing us down, using fourier transform

we end up with only 1 encoder speed is really good.

It is called spectrogram, when you end up converting 1D signal into 2D signal

08/05/2025

2

CLIP

Contrastive learning, very import
used in computer vision

Masked auto encoding