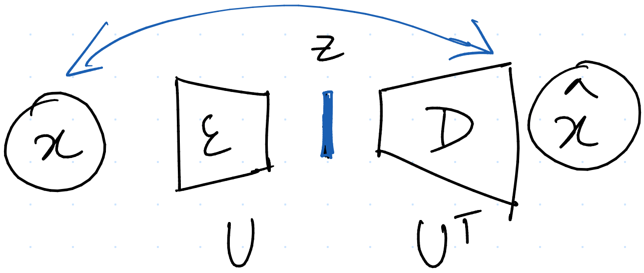
Variational Autoencoders. 8/3/26

Autoencoders. (A.E.'s)



→ meaningful information captured in the latentspace.

$$z = U^T x.$$

$$\hat{x} = Uz$$

$$= U \cdot U^T x.$$

Objective function / loss function or Criterion.

$$\min \| \hat{x} - U U^T x \|^2$$

$$\text{s.t. } U U^T = I.$$

with one layer neurons and without any non-linearity the A.E. behaves as PCA.

There are different types of autoencoders.

$x \in \mathbb{R}^m$

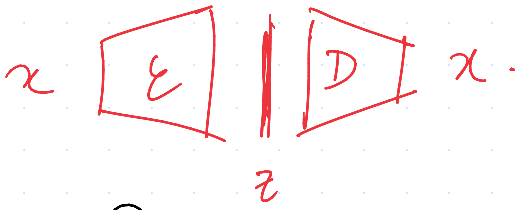
$z \in \mathbb{R}^n$

$m \gg n$.

$x \; \boxed{E} \; z \; \boxed{D} \; \hat{x} \simeq x.$

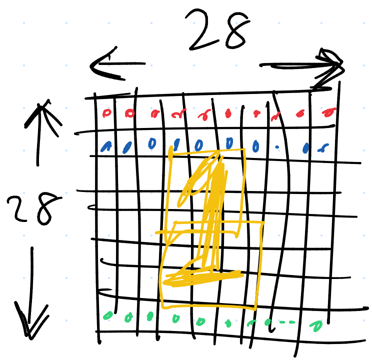
→ o Undercomplete autoencoders.

Overcomplete Autoencoder.



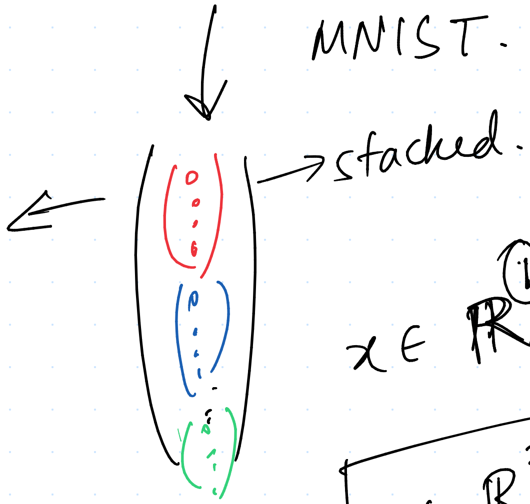
$x \in \mathbb{R}^{m}$

$z \in \mathbb{R}^{n}$ $\therefore$ $n >> m$.



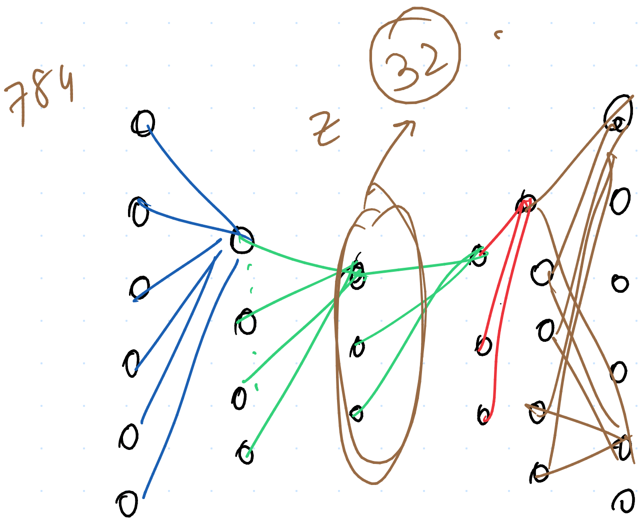
CNN

MNIST.



$x \in \mathbb{R}^{1 \times 784}$

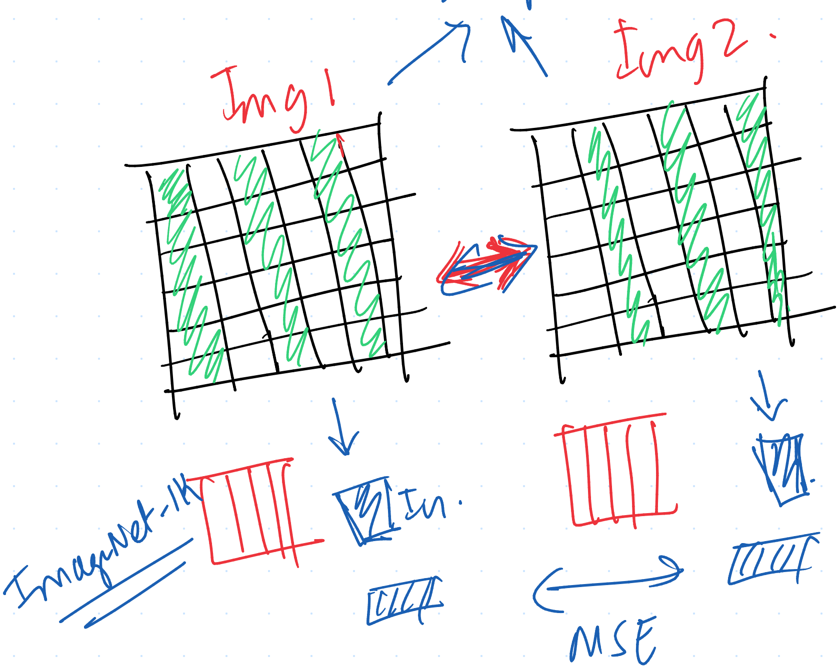
$\boxed{x \in \mathbb{R}^{784}}$



Using fully connected layers.

$\|M(x) - x\|_2^2$

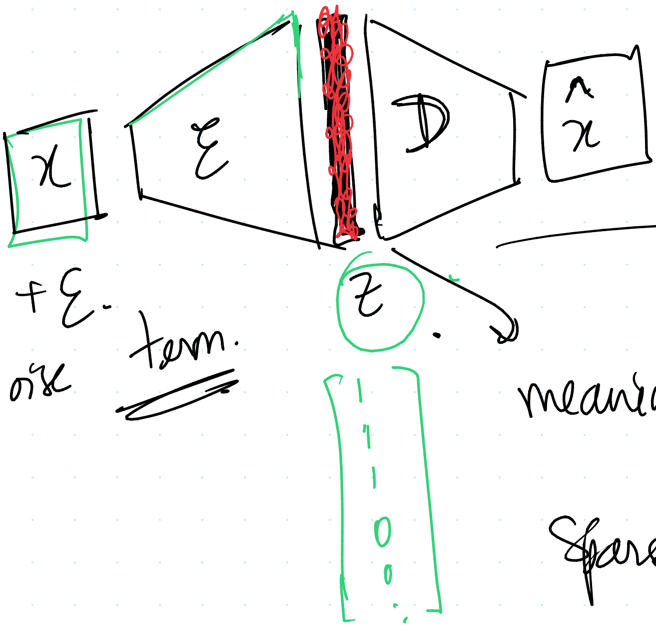
Inception distance.



MSE

→ Large MSE.

- blurry artifacts.
- edgy artifacts.



Identity Mapping

+ε. Noise term.

meaningful representation

Sparse autoencoders.

LLMs → TBs

150B. params. → good interpolation.

Variational AutoEncoders → Stochastic.

z ~ comes from a certain distribution.

$z \sim N(0, I)$. some gaussian (say)

Variational Inference.

$p(z|x)$ → latent variable given an input.

params

z hidden variable

↓

x observed variable.

Probabilistic Graphical Models → representation.

- Topic Modelling (spam/non-spam).
- classification
- Encoding to lower dimension.

Decoder (Model)

likelihood. prior.

posterior..

$$p(z|x) = \frac{p(z,x)}{p(x)} = \frac{p(x|z)\cdot p(z)}{p(x)}$$

→ latent variable.

encoder (Model)

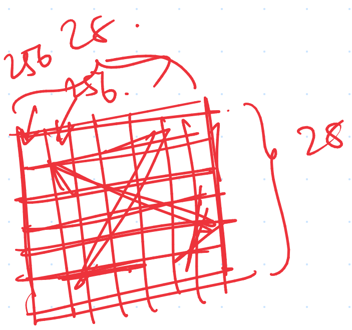
evidence.

intractable quantity

x [E] (z) [D] x

$p(z|x)$. → encoder.

$p(x|z)$ Decoder.



$(256)^{28\times 28} \rightarrow 10^{1889} \simeq 10^{940}$

10^5 ⇒ $0.000\ldots 1\%$

1884→0's

MNIST-sample?

70K → 100 000 →

Grayscale → ▨ (0-255).

Variational Inference

Turn this intractable quantity to an optimization problem, by assuming there is another distribution which is tractable. Now, find the parameters of that distribution that is very close to this one. That distribution is used as a surrogate to the current intractable distribution.

$q_\theta(z) \leftarrow$ comes from a well behaved family of distributions (e.g. Gaussian).

$\min \ KL\left(q(z) \,\|\, p(z|x)\right) \leftarrow$ Minimize the KL divergence b/w these two distributions.

$q(z)$: known.

$p(z|x)$: unknown.

$\simeq q_\theta(z|x) \simeq q_\theta(z)$. Since θ is dependent on x.

$q_\theta(z|x)$: $x \in X$

$q_\theta(z)$: $G(0, I)$.

Information.

$$x \to \boxed{\mathcal{E}} \xrightarrow{z} \boxed{\mathcal{D}} \to x.$$

$\mathcal{E} \sim p(z|x)$ $\qquad$ $\mathcal{D} \sim p(x|z).$

$q_\phi(z|x) \sim q(z).$

$z \sim G(0, I).$

$$I = -\log(p(x)).$$

$x \to$ event.

$p(x) \to$ probability of that event occuring (not occuring)

Higher probability means lower information.

new information.

Mon $\quad$ Tues $\quad$ Wed $\quad$ Thurs $\quad$ Friday $\quad$ Sat.

$p(S) = 0.9$ $\quad$ $p(S) = 0.9$ $\quad$ $p(S) = 0.9$

$p(S) = 0.3$ $\to$ Headlines during drought

Entropy $\to$ Expectation of the information.

$$H = E[I] = -\sum p(x) \log(p(x)) \qquad \boxed{E[x] = \sum x\, p(x)}$$

KL-divergence → more like : Entropy of p − Entropy of q

$$KL(p \| q) = -\sum p(x) \log p(x) + \sum q(x) \log q(x)$$

But in KL we compute the expectation w.r.t. certain quantities, like eg., if the expectation is w.r.t. q, then it is KL divergence.

$$KL(q \| p) = -\sum q(x) \log p(x) + \sum q(x) \log q(x).$$

$$\boxed{KL(q \| p) = -\sum q(x) \log \frac{p(x)}{q(x)}}$$

KL can also be written as the information loss if we want to transfer from one distribution to another, hence, this is a measure b/w two distributions.

Property

- $KL(p \| q) \neq KL(q \| p)$ → hence divergence & not distance
- $KL(p \| q)$ or $KL(\circ \| \circ) \geq 0$

Hence KL is the measure of dissimilarity b/w the two distributions.

So we are minimizing the KL divergence b/w $q_\theta(z)$ & $p(z|x)$ → intractable. Here $q_\theta(z)$ → is from a family of well behaved distribution.

$$\min_\theta \underbrace{KL\left(q_\theta(z) \,\|\, p(z|x)\right)}$$ → continuous form.

$$\Rightarrow -\int_z q(z) \log \frac{p(z|x)}{q(z)} \, dz.$$

$$\Rightarrow -\int_z q(z) \log \frac{p(z,x)}{q(z) \cdot p(x)} \, dz$$

$$\Rightarrow -\int_z q(z) \log \left\{ \underbrace{\frac{p(z,x)}{q(x)} \cdot \frac{1}{p(x)}} \right\} dz$$

$$\Rightarrow -\int_z q(z) \log \frac{p(z,x)}{q(z)} \, dz + \int_z q(z) \log p(x) \, dz$$

$$\log(A \cdot B) = \log A + \log B.$$

$$\Rightarrow -\int_z q(z) \log \frac{p(z,x)}{q(z)} dz + \underbrace{\log p(x)}\ \underbrace{\int_z q(z)\, dz}_{1.}$$

Since we are integrating on z and $p(x)$ observation which is a constant, and it doesn't depends on z or anything, hence we are taking it out $\rightarrow$ Nothing to do with θ either.

$$\underline{\min}\ KL\left(q(z) \| p(z|x)\right) = -\int q(z) \log \frac{p(z,x)}{q(z)} + \underbrace{\log p(x)}_{\text{constant, tractable}}$$

maximize this quantity

from dataset.

ELBO - Evidence Lower Bound.

VLB - Variational Lower Bound.

tractable

known → dataset

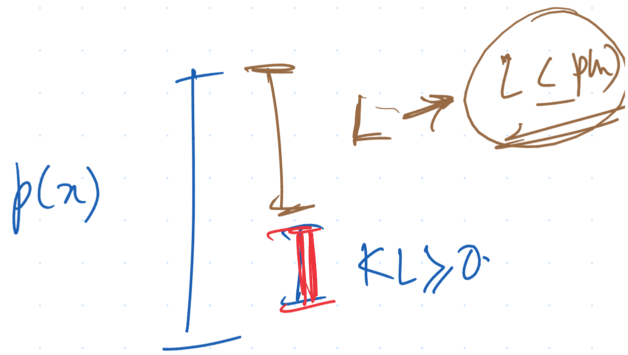
$$q_\theta(z) \cdot \log \frac{p(z,x)}{q_\theta(z)} \longrightarrow \text{Encoder learnt representation} \rightarrow \text{tractable}$$

tractable

$$\underbrace{\log p(x)}_{\text{Constant}} = \underbrace{KL\left(q(z) \,\|\, p(z|x)\right)}_{\geq 0} + \underbrace{\int q(z) \log \frac{p(z,x)}{q(z)}}_{\mathcal{L} \text{ or lower bound}}$$

$\mathcal{L} \neq \log p(x)$ unless $KL = 0$.

Hence, $\mathcal{L}$ is the lower bound of the $p(x)$.



Lower Bound.

$$p(x|z) = \frac{p(x,z)}{p(z)}$$

$$p(x,z) = p(x|z) \cdot p(z)$$

$$\mathcal{L} = \underline{\int q(z) \log \frac{p(z,x)}{q(z)}}$$

Decoder.

$$= \int q(z) \log \frac{p(x|z) \cdot p(z)}{q(z)}$$

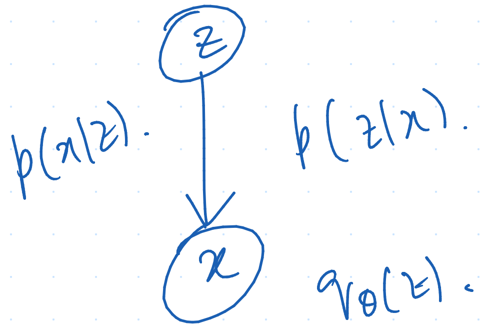
well defined Gaussian.

$$= \underbrace{\int q(z) \log p(x|z)}_{\text{maximize } \mathcal{L}} + \underbrace{\int q_\theta(z) \log \frac{p(z)}{q_\theta(z)}}_{-KL(q(z)\,\|\,p(z))}.$$

$$= \max \int q(z) \log p(x|z)$$

$$= \max E[\log p(x|z)] \rightarrow \text{likelihood of the dataset.}$$

$$\log p(x) = KL\left(q(z) \,\|\, p(z|x)\right) + E\left[p(x|z)\right] - KL\left(q(z) \,\|\, p(z)\right)$$



<u>Maximize likelihood:-</u>

→ Gaussian: minimize MSE

→ Bernoulli – minimize CE loss.

Gaussian

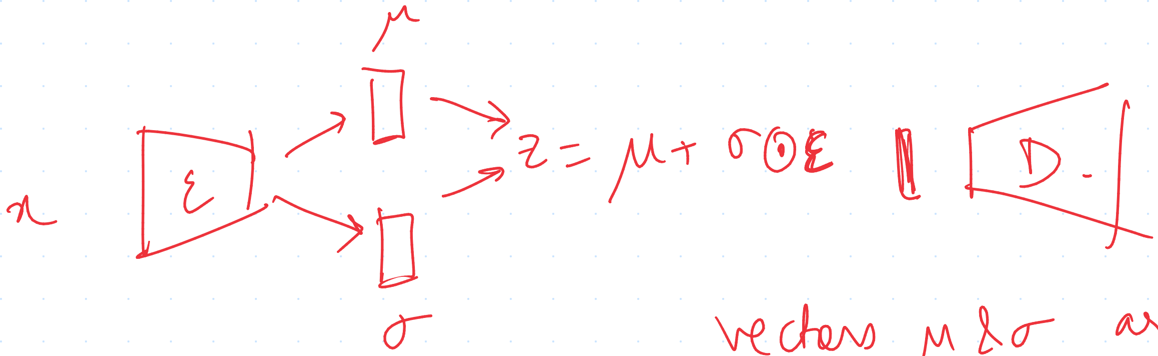
$$\underbrace{|x - \hat{x}|}_{AE} + KL\underbrace{\left(q(z) \,\|\, N(0,0)\right)}_{VAE}$$

This additional loss in VAEs ensures that the z is Gaussian, since z → stochastic, hence no backpropagation.

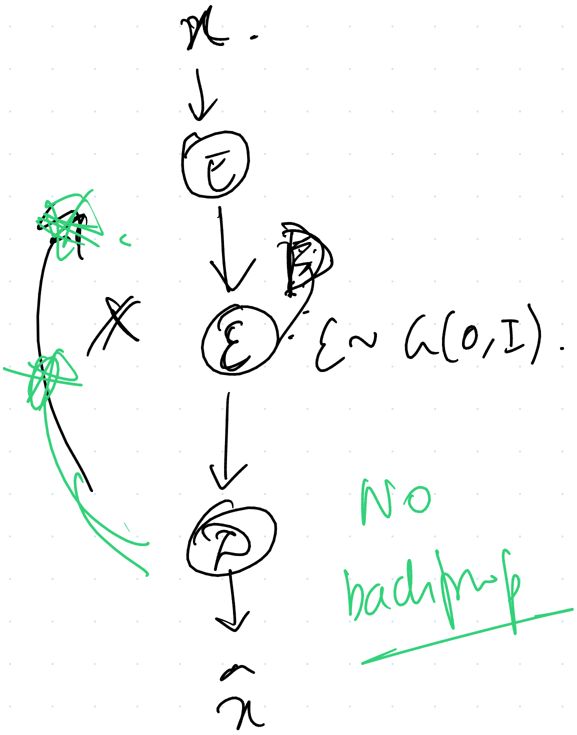
Reparameterization trick :- Find the mean & variance of the dist via the neural network.

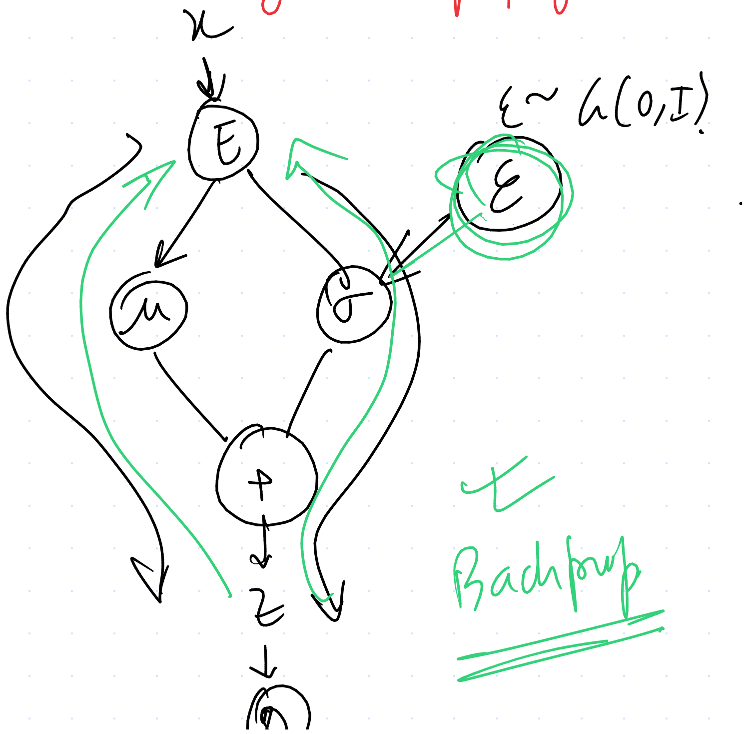
(mean, variance) → deterministic.

Through this Gaussian - sample something random, representation of z → parameters of z in the model.



vectors μ & σ are learnt using backpropagation.





$\tilde{x}$

- Inpainting.
- Grayscale → colored images (colorization).