VAEs (Variational Auto-Encoders)

26/02/2025

Autoencoders. (A.E's)

$x \rightarrow \boxed{E} \rightarrow z \rightarrow \boxed{D} \rightarrow \hat{x}$

$U \qquad U^T$

$z = U^T x$

$\hat{x} = U z$

$= U U^T x.$

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

Objective function :-

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

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

$I I^T = I$

Autoencoders → used for compression. → undercomplete autoencoder.

$x \in \mathbb{R}^m \qquad m >> n.$

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

$x \rightarrow \boxed{E} \rightarrow z \rightarrow \boxed{D} \rightarrow \hat{x} \approx x.$

undercomplete autoencoder.

$x \rightarrow \boxed{E} \rightarrow \boxed{D} \rightarrow x.$

$x \in \mathbb{R}^m \qquad n >> m.$

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

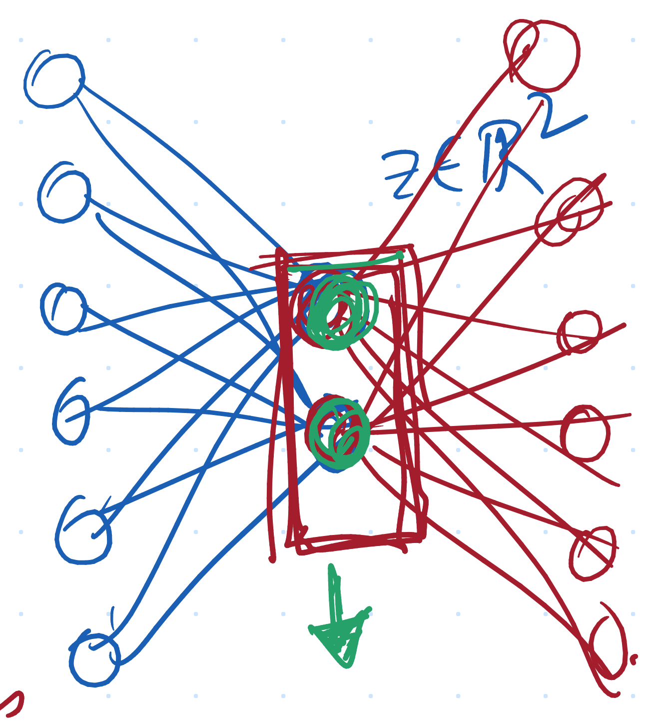
$x \in \mathbb{R}^6$

$x.$

$x \in \mathbb{R}^6.$

$z \in \mathbb{R}^2$

$\tilde{x}$



redundant features.

$\mathbb{R}^1$

28

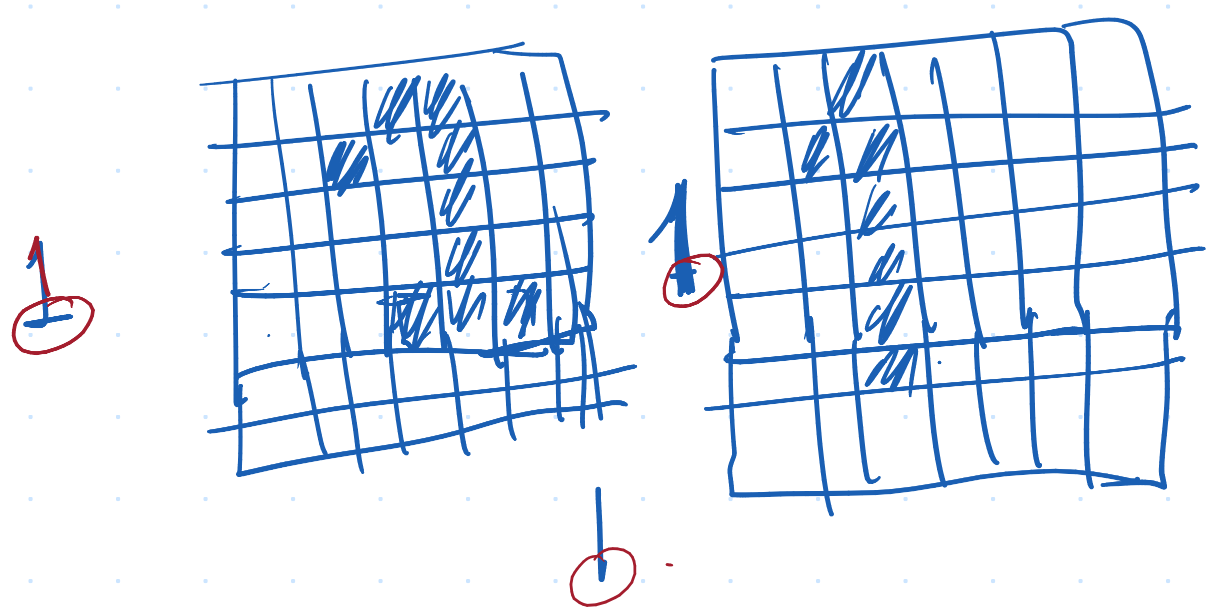
28

x^1

→

$y^1 \in \{0, 1, 2 .. 9\}.$

7 X 7



$$\begin{array}{r} 28 \\ \times 28 \\ \hline 224 \\ 56\times \\ \hline 784 \end{array}$$

$28 \times 28 \Rightarrow 784.$

$x \in \mathbb{R}^{m \times m}$ higher dimensional.

$z \in \mathbb{R}^{n}$ where $m >> n$, is always lower dimensional.

→ no labels.

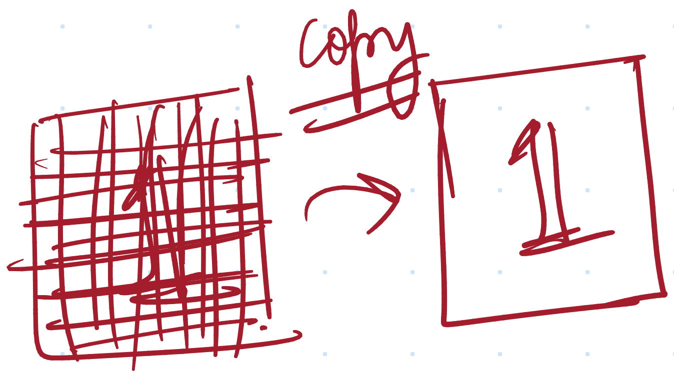
MSE loss.

L1 loss.

L2/L1 minimize. for similarity

$(5, 7, 3, 2, 1)$

$(5,7,5,2,2)$ $(5, 7, 3, 2, 1)$ $(5,7,1,2,1)$



$$\cos\theta = \frac{a \cdot b}{\|a\| \|b\|} = \frac{25+49+9+4+1}{\left(\sqrt{25+49+9+4+1}\right)^2} = \frac{\cancel{1}}{\cancel{1}} = 1$$

$$\boxed{\theta = \cos^{-1}(1) = 0}.$$

$\underbrace{[-1 \text{ to } 1]}$.

$(1,2,3)$ $(4,5,6)$

$$\Rightarrow \sqrt{(4-1)^2 + (5-2)^2 + (6-3)^2} \approx \text{high}.$$

but cosine similarity $= 0$.

$A = [1, 0]$ $B = [-1, 0]$.

$x = [3, 4]$ $\hat{x} = [0; 0]$.

$\theta = 90$

Cosine

$(-1, 1)$.

MSE. ✗.

$(0, \infty)$

L1

$(0, \infty)$.

$x = (3, 4)$

$\hat{x} = (-4, -3)$

$$C.S. = \frac{x \cdot \hat{x}}{\|x\| \|\hat{x}\|} = \frac{-24}{25} = -0.96$$

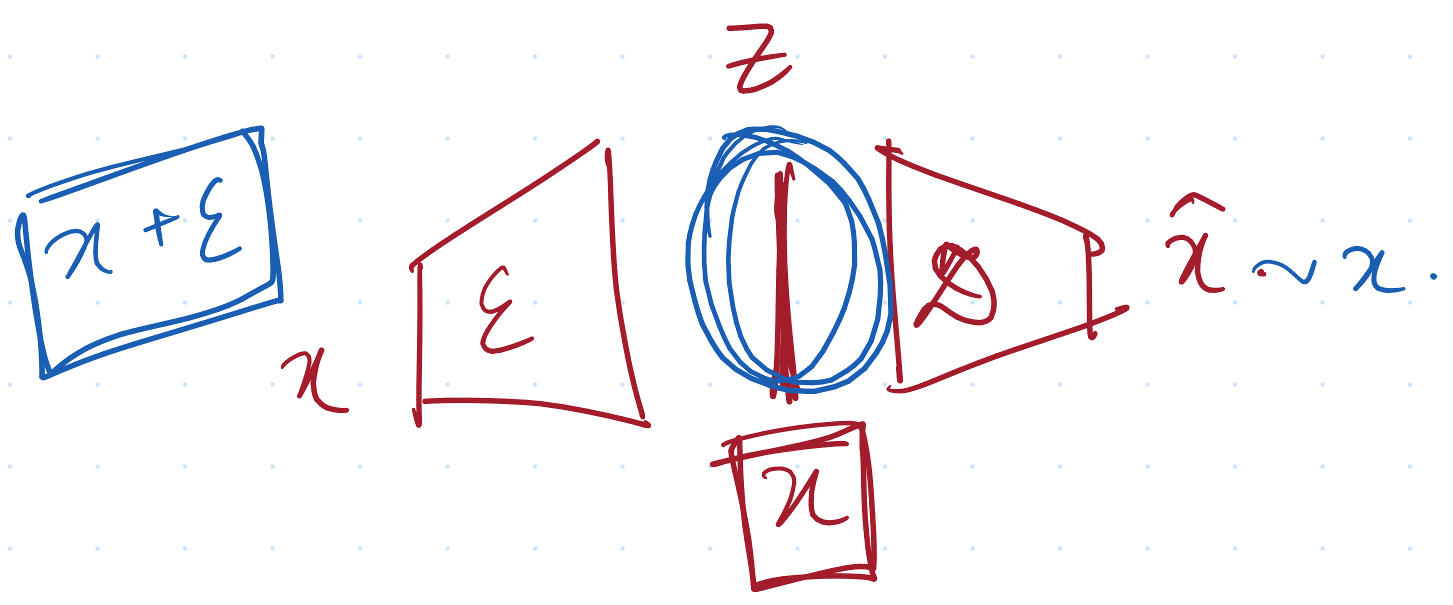
$\sqrt{(-4)^2 + (-3)^2}$

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (x_i - \hat{x}_i)^2$$

$$= \frac{1}{2} \left((3 - (-4))^2 + (4 - (-3))^2 \right) = \frac{49 + 49}{2}$$

$$= 49.$$

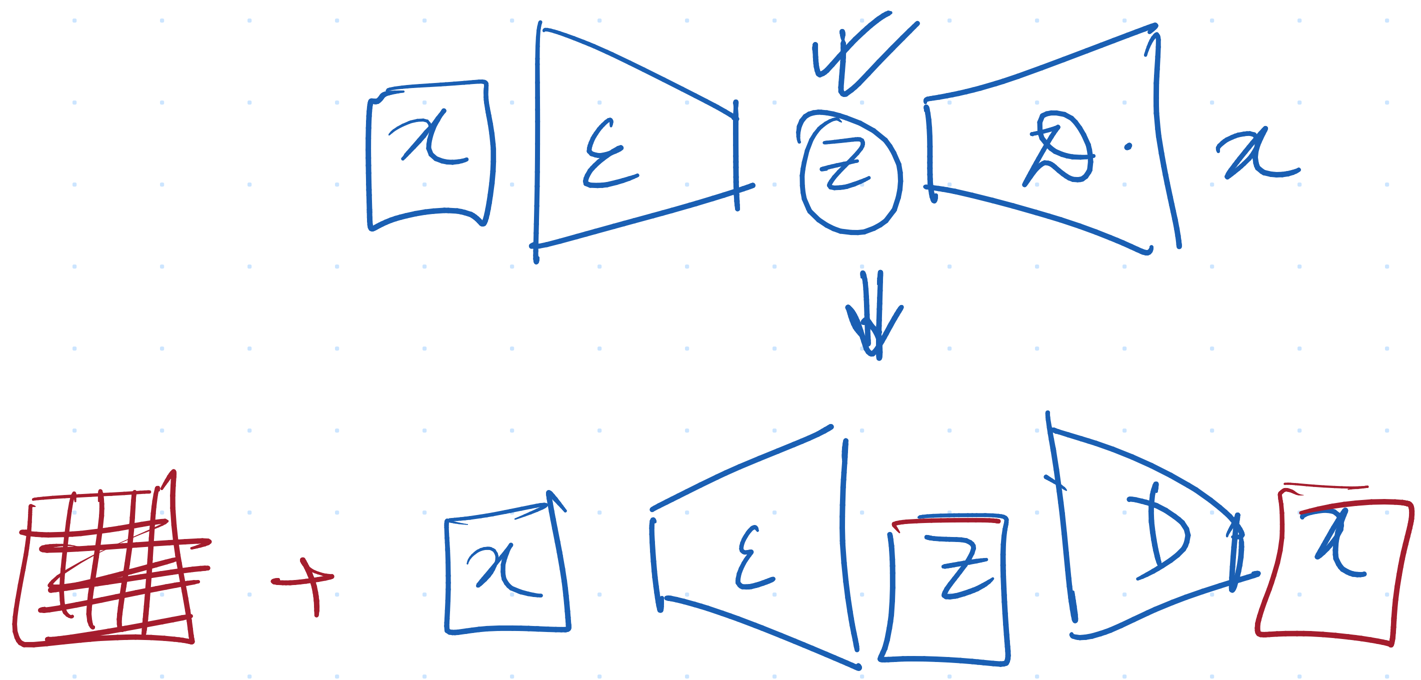
MSE, L1 good for better representation.



$z > x$.

$\| x - \hat{x} \|_2$

identity mapping is being learnt.



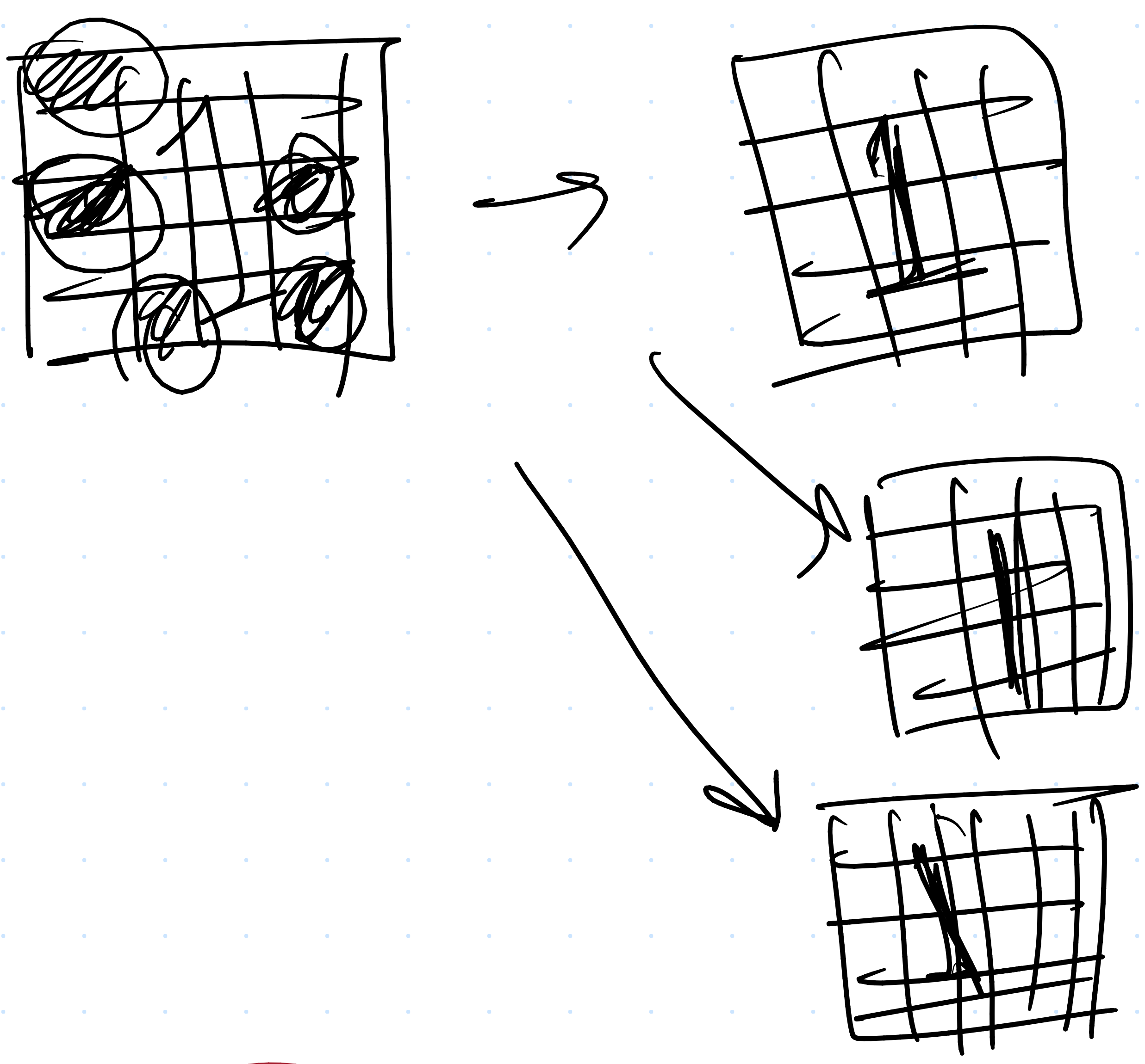
Noise Removal.

Denoising autoencoder.

$\boxed{x+\varepsilon}$

identity mapping

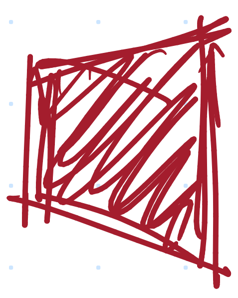
$U = I$.



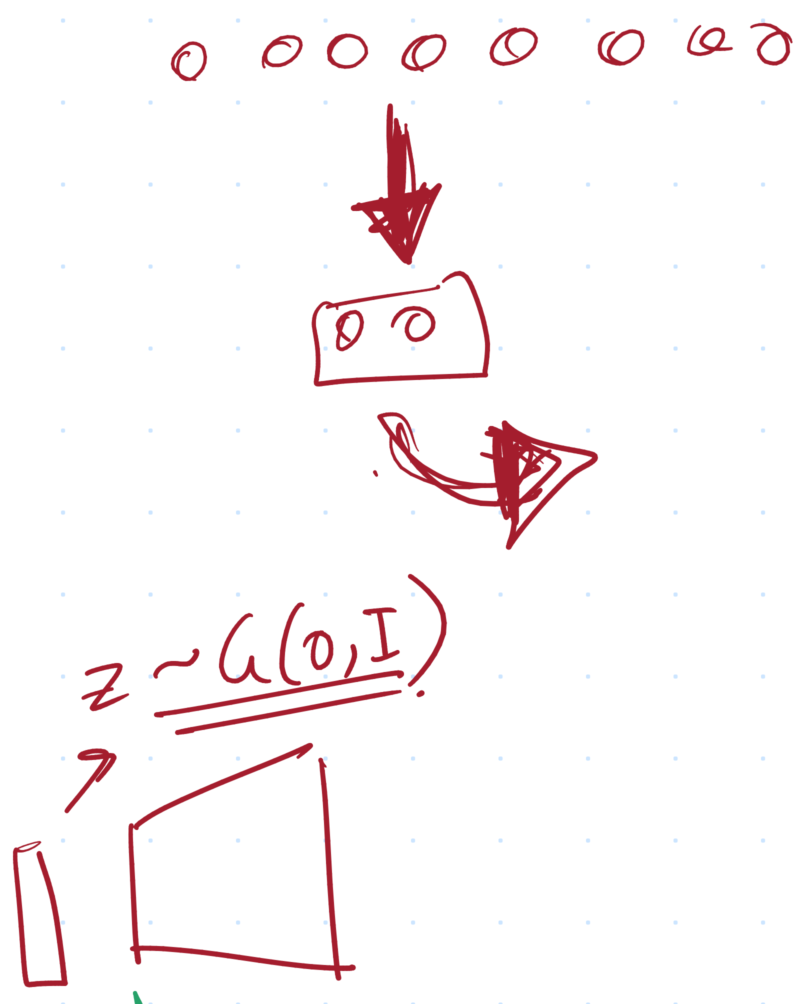
100%

outliers.

95% → Good image + Noise.



Variational AE - stochastic.



$z \to$ comes from a certain distribution.

$z \sim G(0,I)$ some Gaussian, say.

Variational Inference $\to$ Bayesian statistics

(z) hidden variable

$z \to x$

$p(z|x) \to$ latent var. given an input.

- Topic Modelling
- classification.
- Encoding to lower dim, etc.

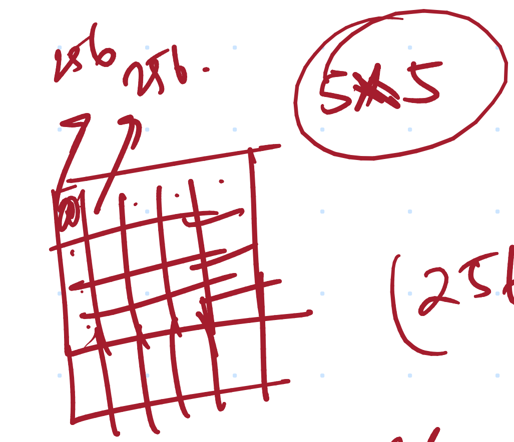
$$p(z|x) = \frac{p(x|z)\,p(z)}{p(x)}$$

$p(x|z) \to$ Decoder/Model.

$p(z) \to$ prior.

$p(x) = \int_z p(x|z) \cdot p(z)\, dz \leftarrow$ Intractable quantity

$\to$ when $z \to$ higher dimension, then $\int\int\int\int \dots$ ($z_1 z_2 z_3 z_4$) this becomes complicated integral.



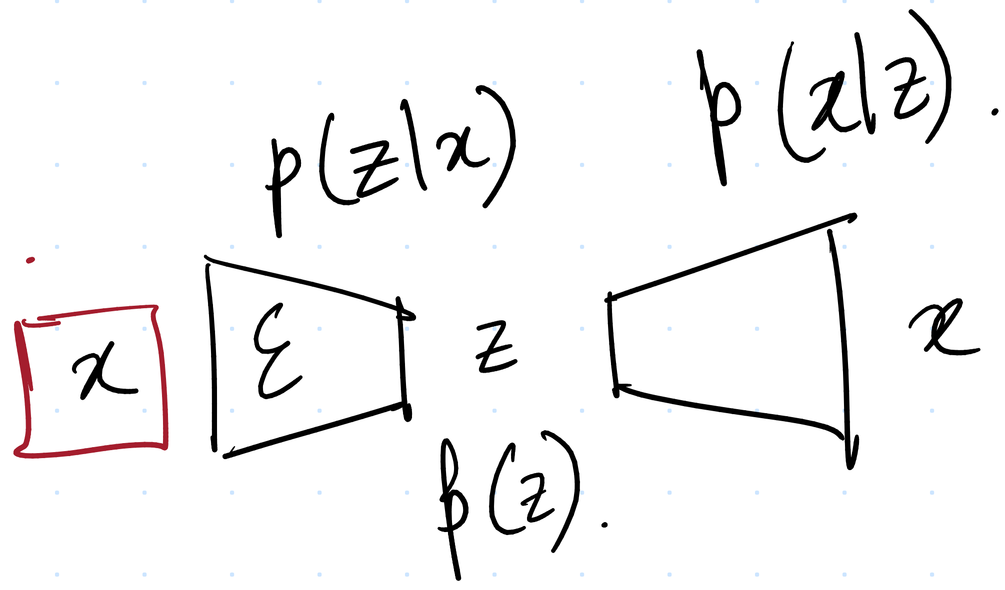
$(256)^{25}$

Grayscale image.

$(0-255) \to 256$

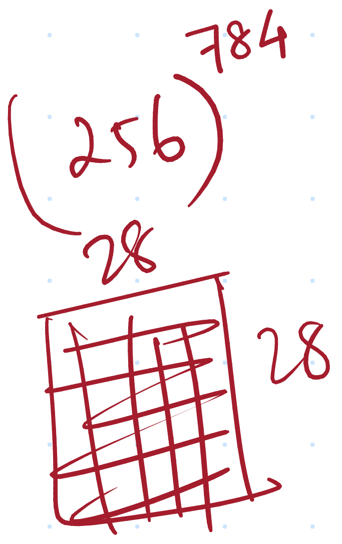
$(256)^{784}$

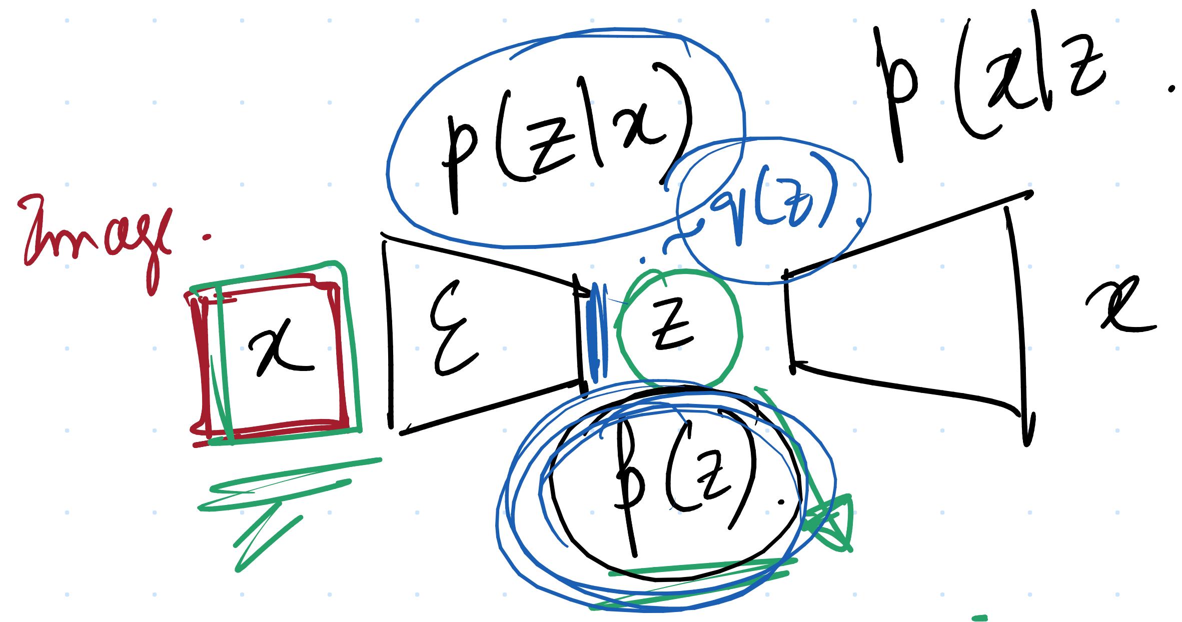
Image.



* Sampling techniques $\to$ one line of technique to address this intractability.

$$p(z|x) = \frac{p(x|z)\, p(z)}{p(x)} \rightarrow \text{prior.}$$





So this is intractable

(1998) 0's

60K

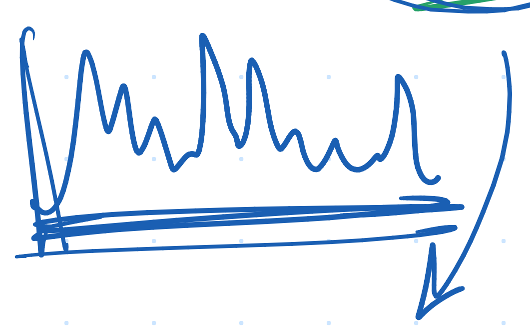
7(1998)

10-1998

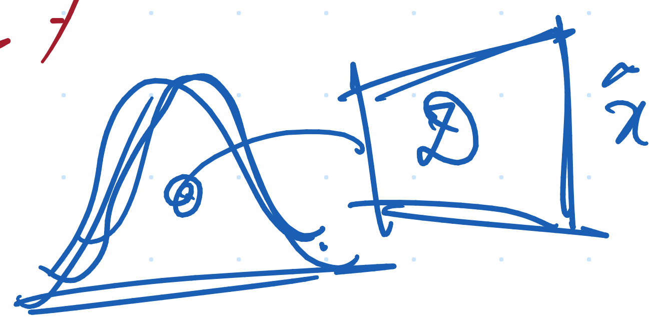
(0.00.... 1 %)

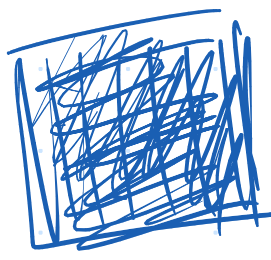
since we are not getting x fully like 0.00 ---- 01% of the real data in even 28x28 pixel which is barely recognizable & even graysale (not even colored)

1998 of 0's



$q_\theta(z) \sim G(0, I)$





→ blurry average.

$\hat{x} \sim x$.

⊛ Variational Infuence

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. (like Gaussian)

$$\min KL\left(q(z) \,\|\, p(z|x)\right)$$

← Minimize the KL divergence b/w these two distributions

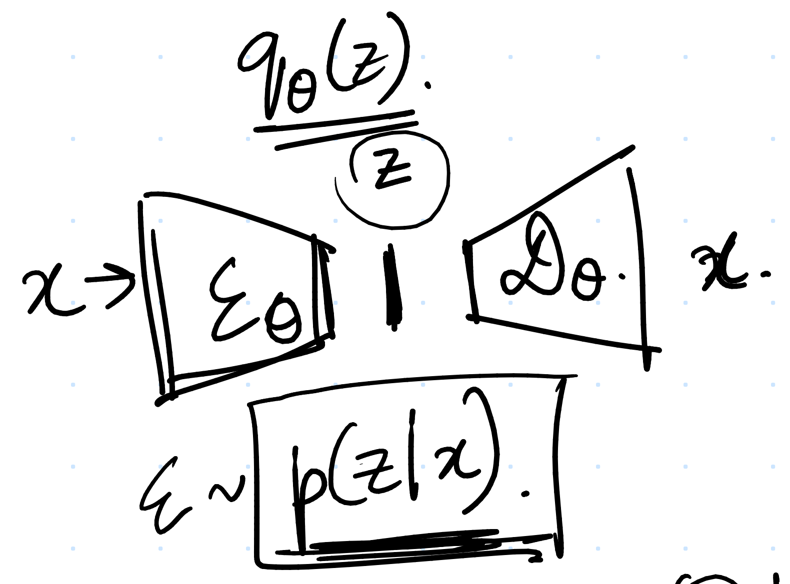
known. unknown.

$\approx q_\theta(z|x) \approx q_\theta(z)$. since θ is dependent on x.

$0.00\cdots 1\%$.

$x \in X$

$\sim \mathcal{G}(0,I)$.



$p(x)$ → actual dist.

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

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

$z \sim \mathcal{G}(0,I)$.

Information

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

$x \rightarrow$ event.

Higher probability means lower information

Entropy → Expectation of the information.

$$E[x] = \sum x\, p(x).$$

$$H = E[I] = -\sum p(x) \log(p(x)).$$

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 this is KL divergence.

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

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

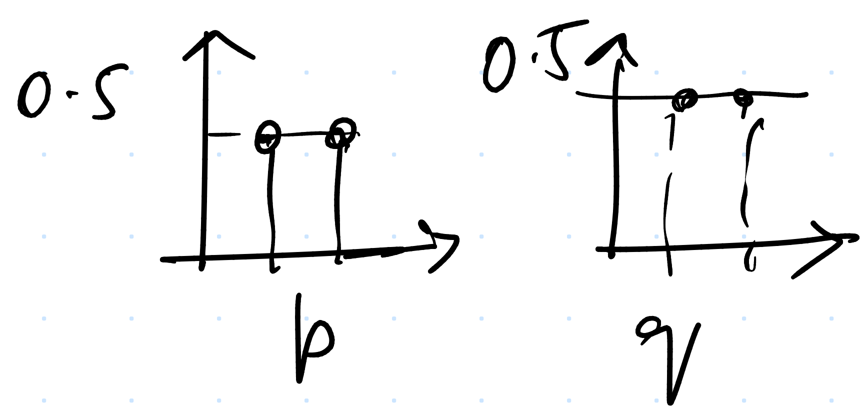
KL can also be written as the information loss if we want to transfer from one dist to another, hence, this is a measure b/w two distributions.

Property of KL divergence

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

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

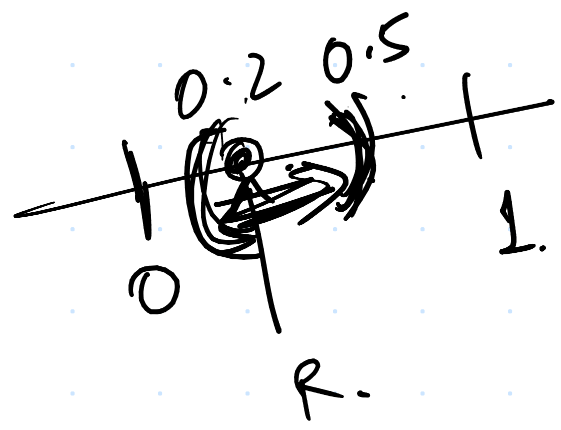
$KL(p\|q)$



$$= -\sum q \log \frac{p(x)}{q(x)}.$$

$$= -\sum q \log \frac{0.5}{0.5} = -\sum q \log(1) = 0$$

So, we were minimizing the KL divergence b/w $q_\theta(z)$ and $p(z|x)$ → intractable. Here $q_\theta(z)$ → well behaved family of distribution.



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

$p(x) = 0.$

$\frac{0.5-0.2}{1-0} \Rightarrow \frac{0.3}{1} = 0.3$

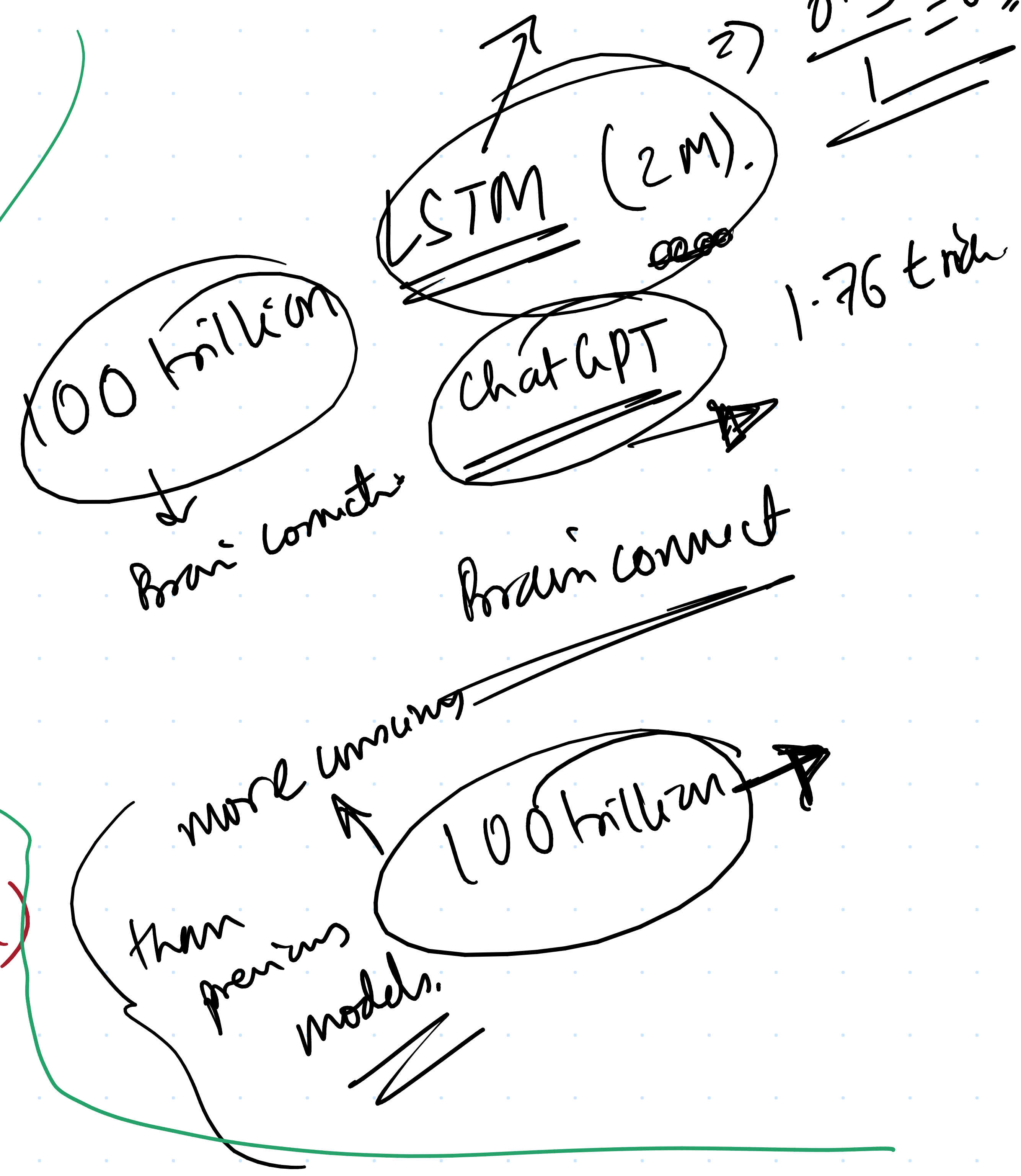
$KL\left(q(z) \,\|\, p(z|x)\right) \rightarrow$ KLD in continuous space

$$= -\int_z q(z) \log \frac{p(z|x)}{q(z)}$$

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

$$= -\int_z q(z) \log \frac{p(z,x)}{q(z)} \cdot \frac{1}{p(x)}$$

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



$$\therefore \min KL\left(q(z) \,\|\, p(z|x)\right) = -\int q(z) \log \frac{p(z,x)}{q(z)} + \log p(x) \int \overbrace{q(z)\,dz}^{1}$$

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. → Nothing to do with θ either.

$$\min \; KL\left(q(z) \,\|\, p(z|x)\right) = -\underbrace{\int q(z) \log \frac{p(z,x)}{q(z)}}_{\text{maximize this quantity.}} + \underbrace{\log p(x)}_{\text{constant. \& tractable.}}$$

ELBO - Evidence Lower Bound.

VLB - Variational lower Bound.

$$\underbrace{q(z)}_{\text{tractable.}} \log \frac{p(z, x)}{q(z)}$$

gaussian → latent

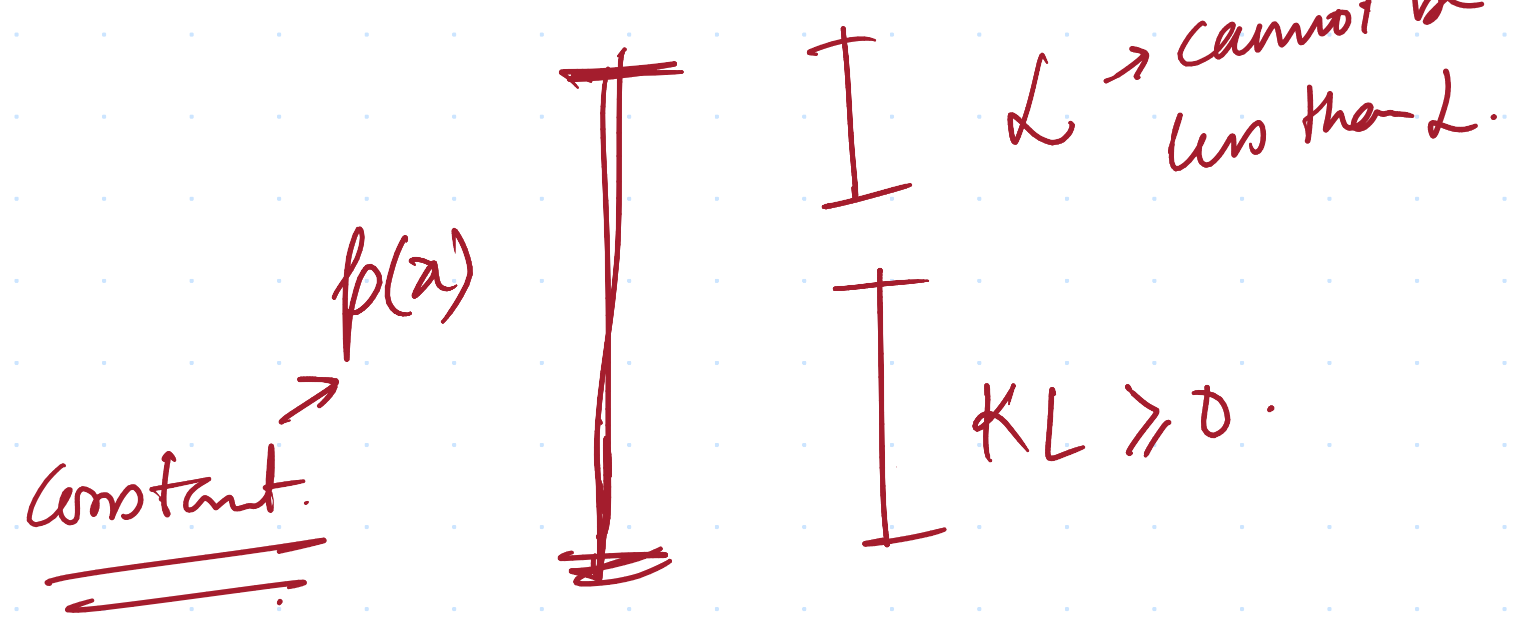
$p(z, x)$ → tractable. → Encoder learnt.

$q(z)$ ← tractable, gaussian

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

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

Hence, $\mathcal{L}$ = lower bond of this $p(x)$.



Lower Bound :-

$q(z)$

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

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

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

Decoder

$$= \int q(z) \log \frac{p(x|z)\, p(z)}{q(z)} \leftarrow \text{well defined Gaussian.}$$

$$= \underbrace{\int q(z) \log p(x|z)}_{} + \underbrace{\int q(z) \log \frac{p(z)}{q(z)}}_{-KL(q\|p)}$$

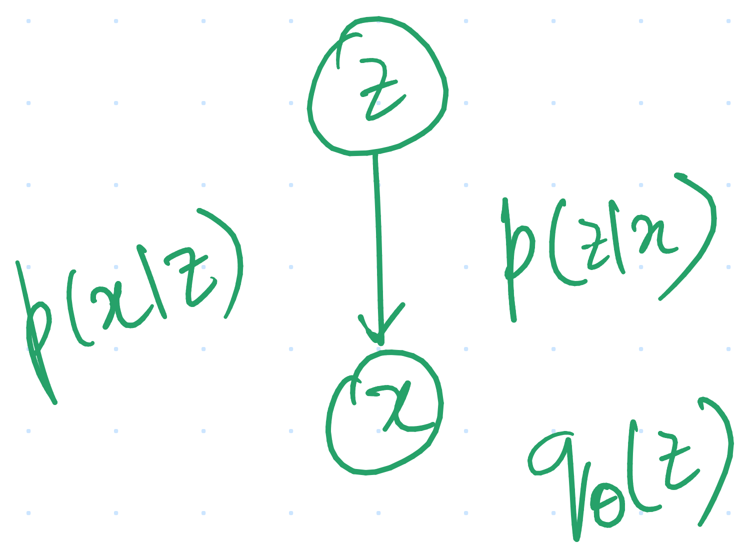
$q(z)$ well behaved gaussian distribution.

maximize $\mathcal{L}$

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

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

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



Maximize Likelihood :-

→ Gaussian - minimize MSE

→ Bernoulli → minimize CE loss.

Gaussian :-

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

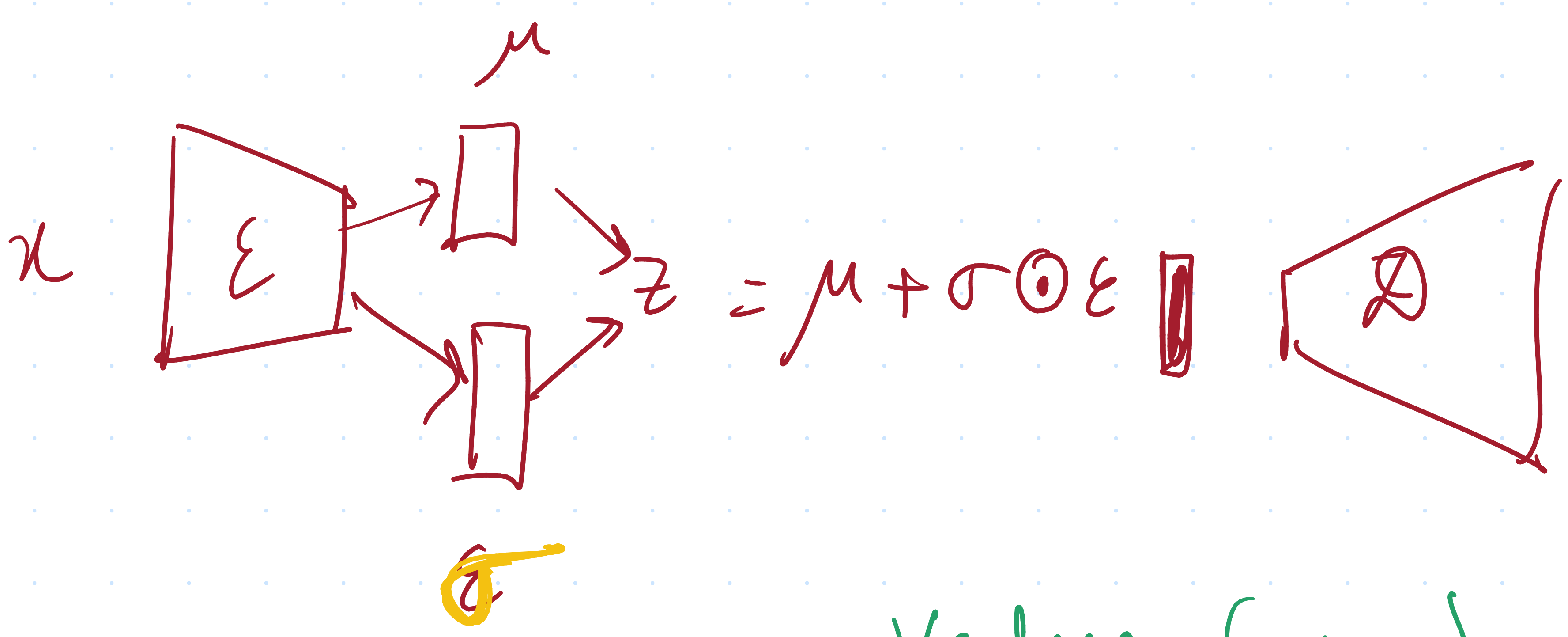
This additional loss in VAEs ensures that the z is Gaussian.

Since $z \to$ stochastic hence no backpropagation.

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

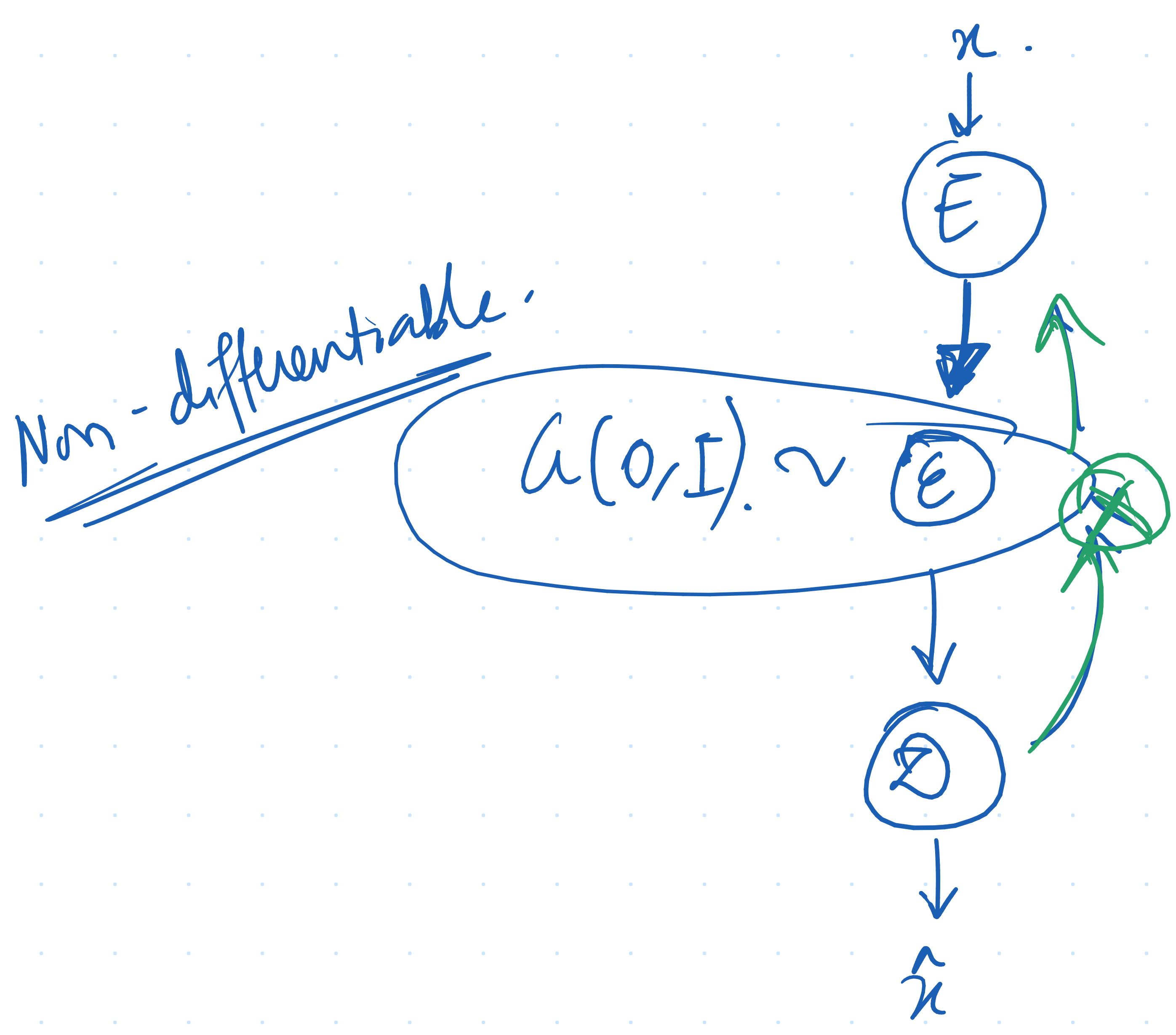
(mean, variance)
↓
deterministic.

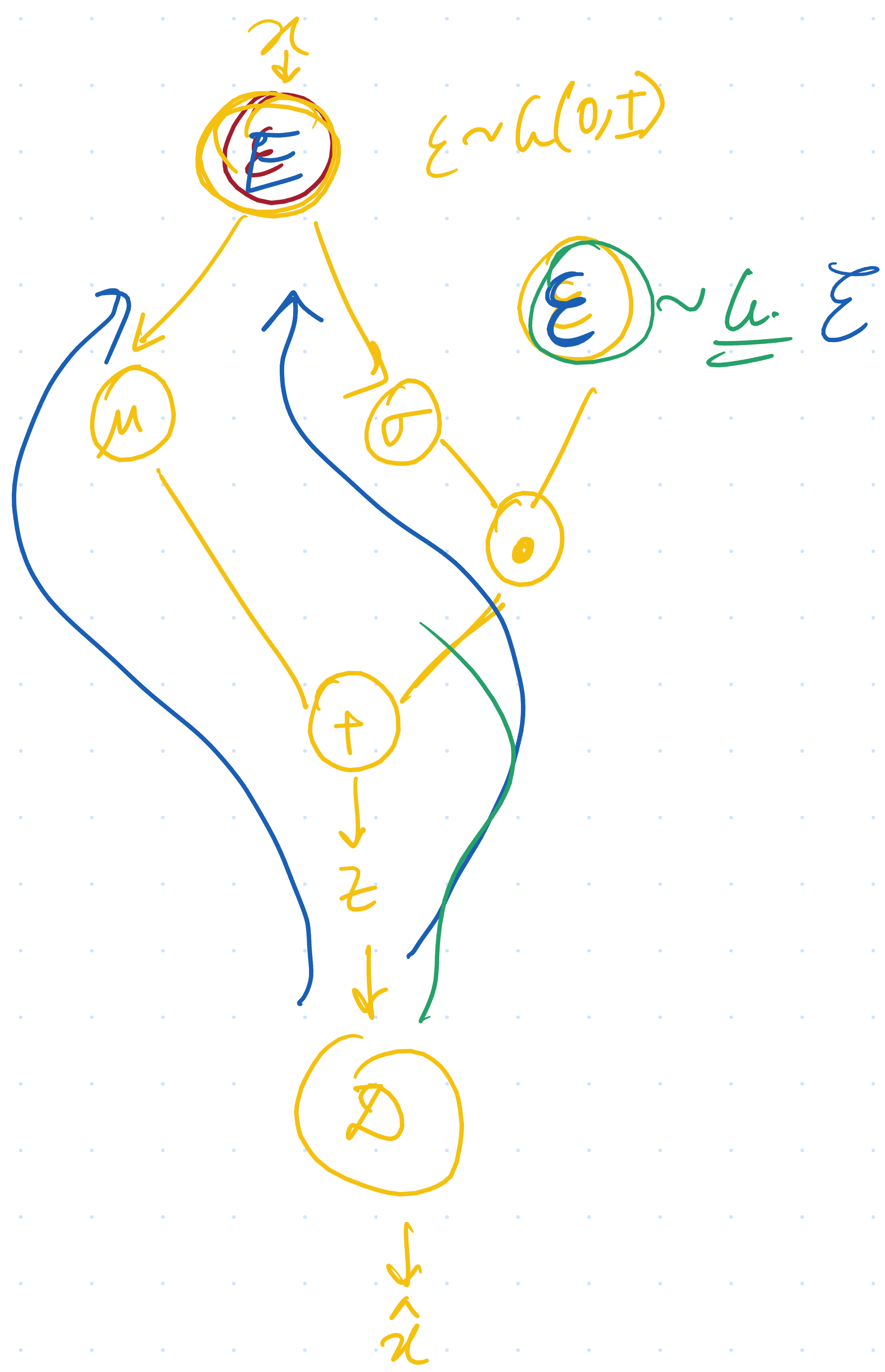
Through this Gaussian → sample something random; representation of z → parameters of z in the model.



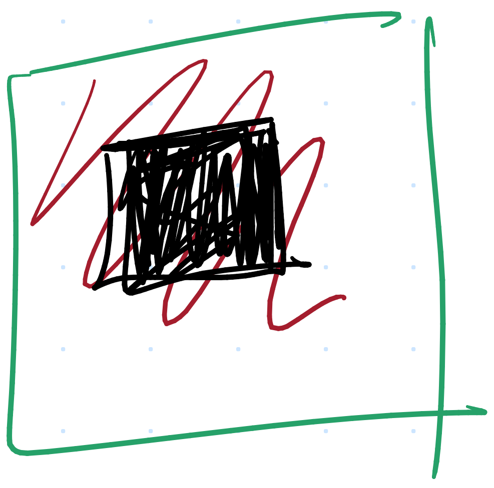
Vectors (μ, σ) are learned by backpropagation.

(Idea of reparameterization)

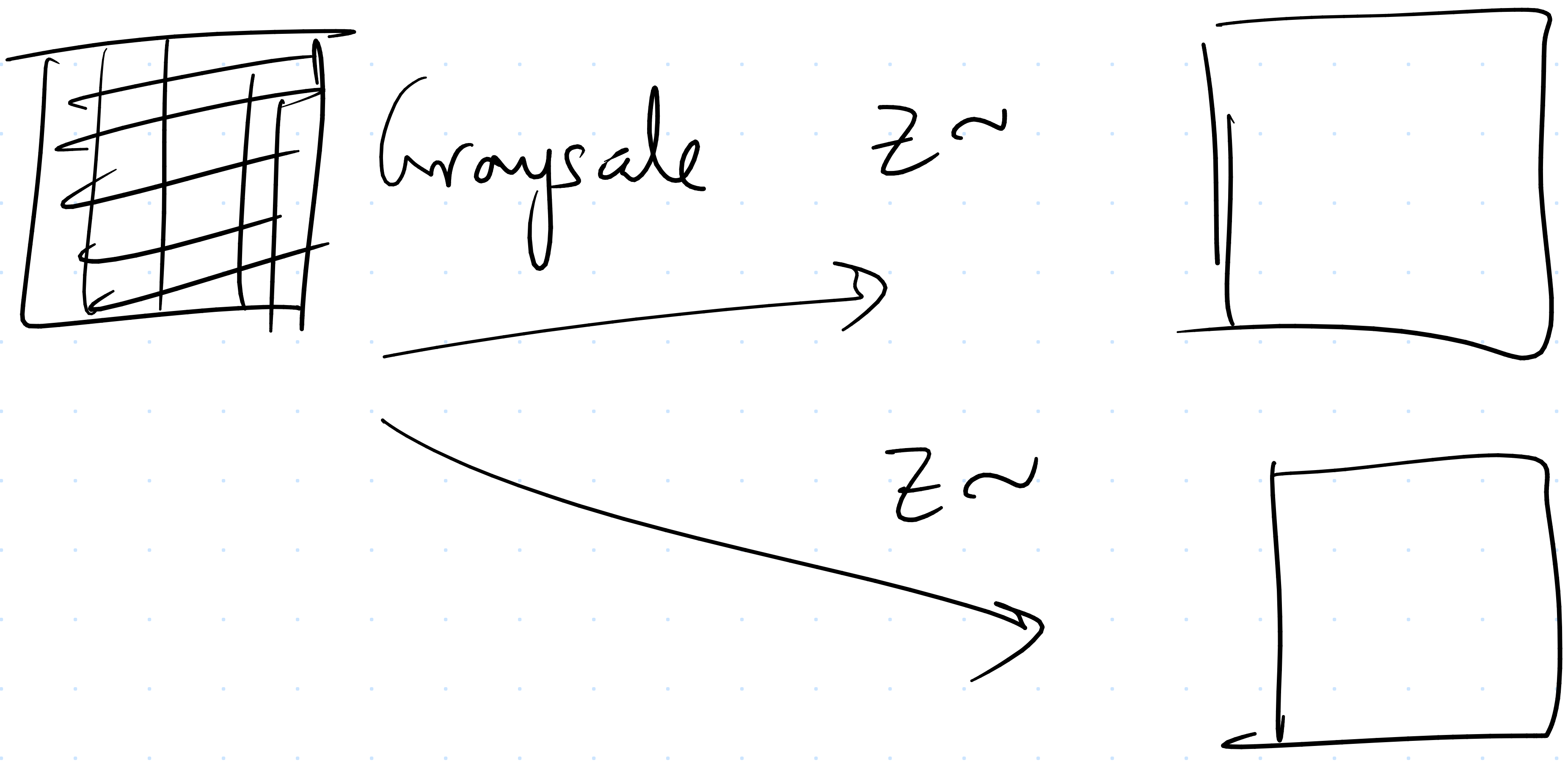




Inpainting



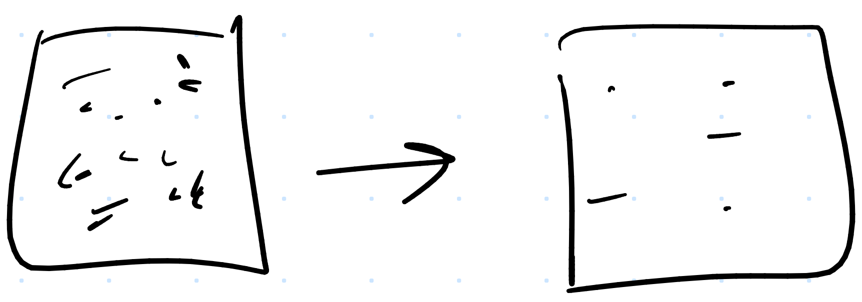
Colorization



Colored image.

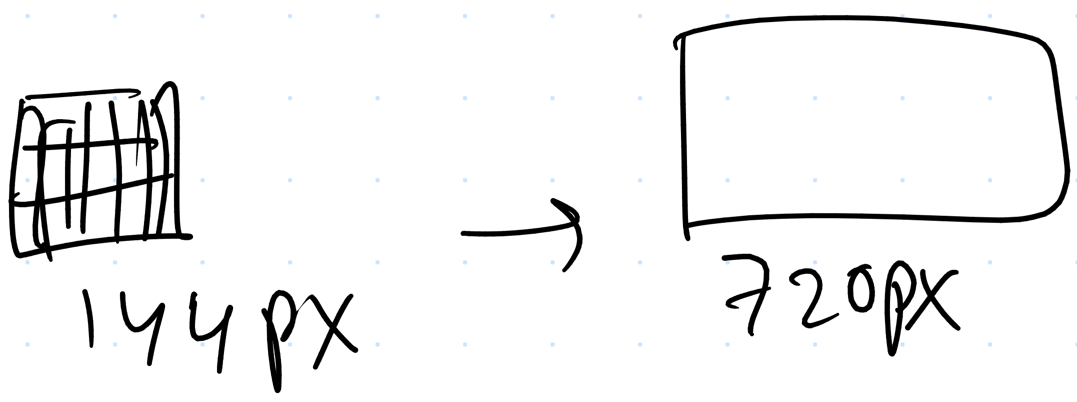
colored.

Old photos →



New photo

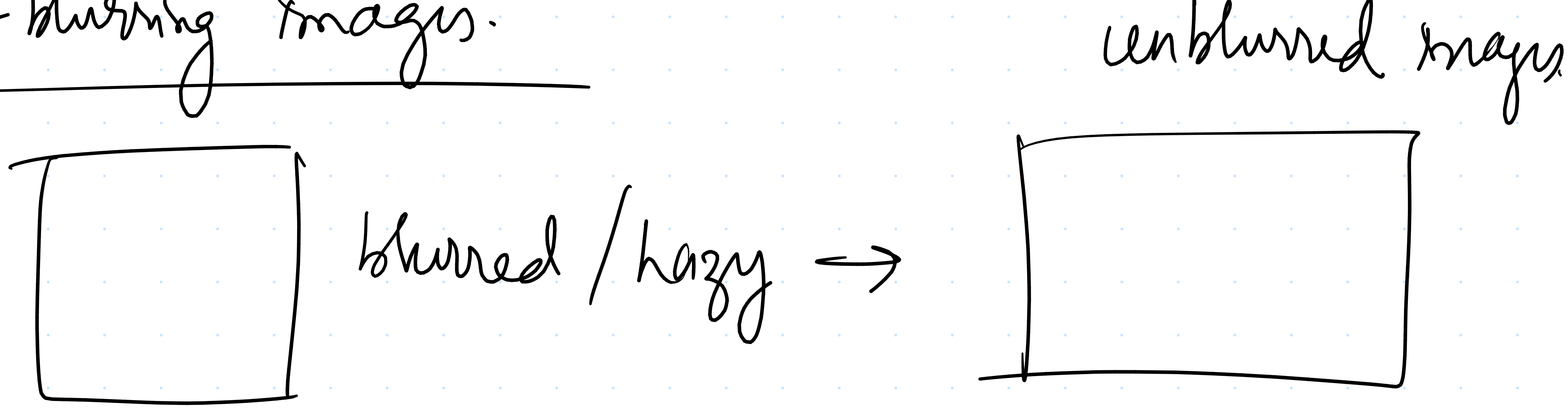
Super-resolution

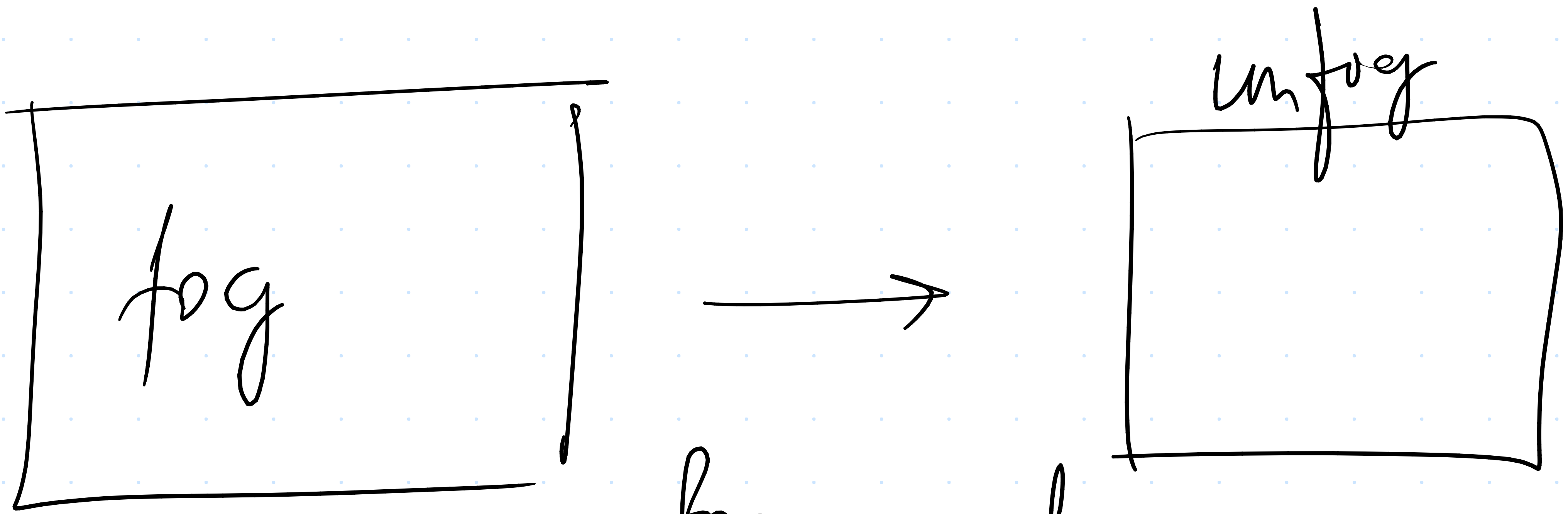


High resolution

Depth map. → realistic images.

De-blurring images.





fog removal.