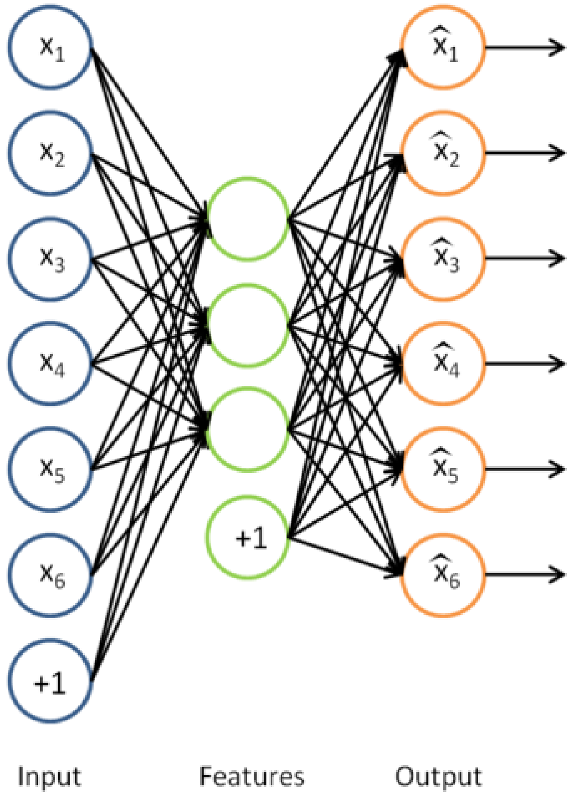
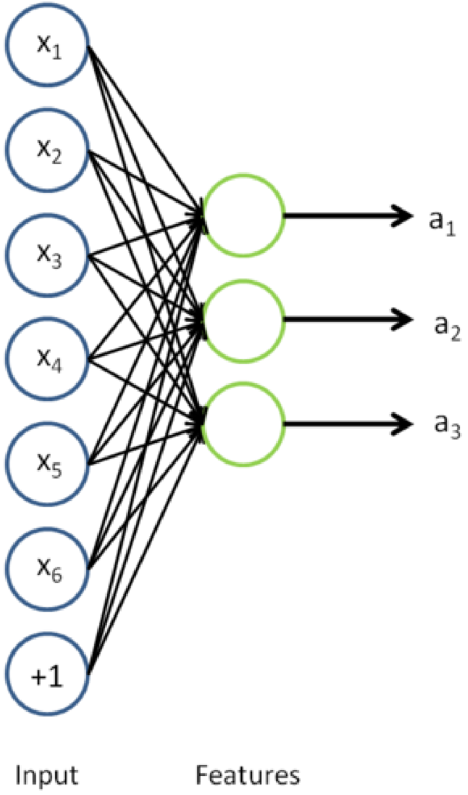
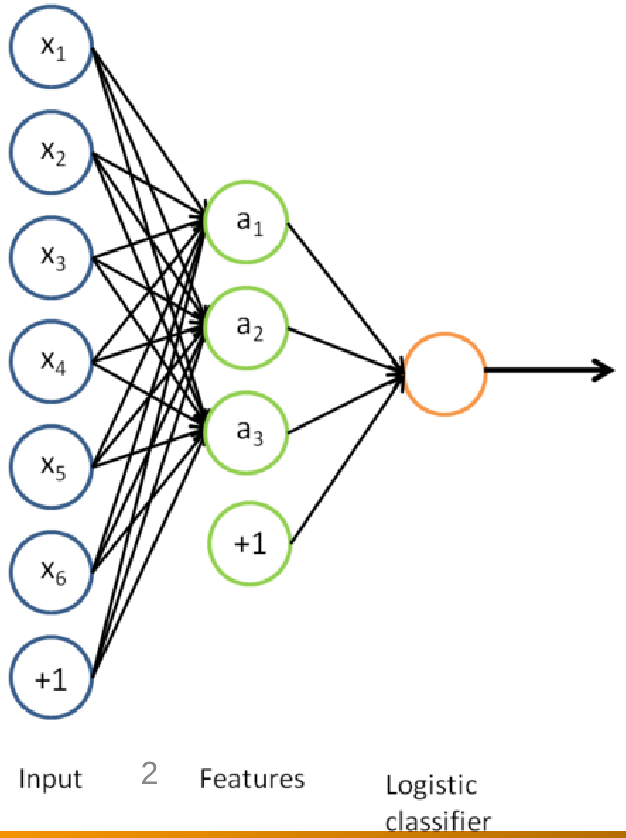
AutoEncoder

Auto-Encoders

- A type of unsupervised learning which tries to discover generic features of the data
  - Learn identity function by learning important sub-features (not by just passing through data)
  - Compression, etc.
  - Can use just new features in the new training set or concatenate both

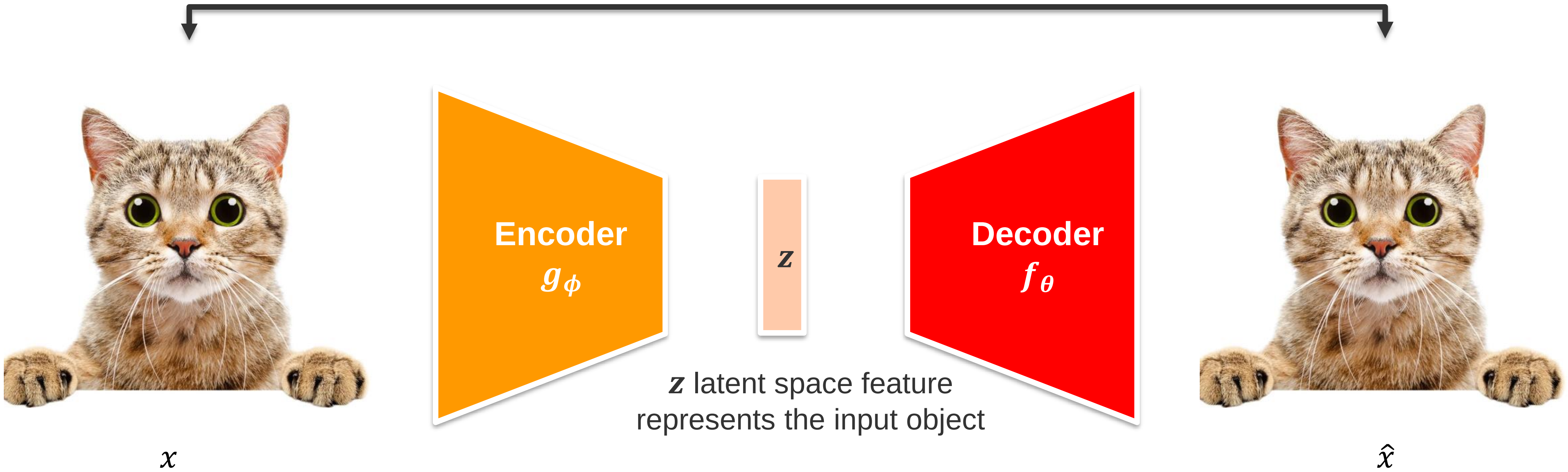






Autoencoder

Reconstruction Loss $$\mathcal{L}(\phi,\theta) = \left(x - f_\theta\left(g_\phi(x)\right)\right)^2$$



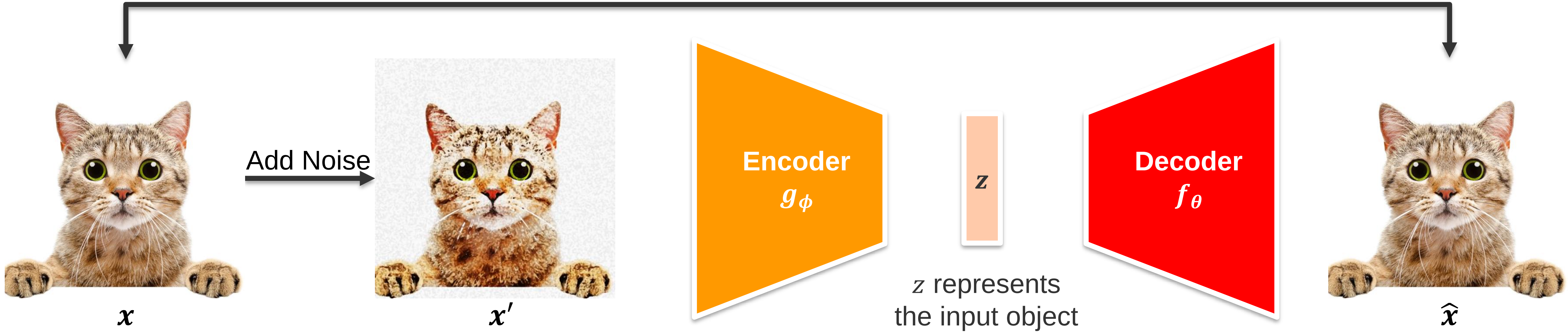
z latent space feature
represents the input object

- The decoder tries to reconstruct the input image based on the encoding of the input
- To reconstruct the image, the encoder should learn good representations of the input
- The self-supervised pre-trained model can then be used to train on the target task

Denoising Autoencoder [2]

Reconstruction Loss $\mathcal{L}(\phi,\theta) = \left(x - f_\theta\left(g_\phi(x')\right)\right)^2$

即便加了雜訊所取出之特徵，仍要能還原

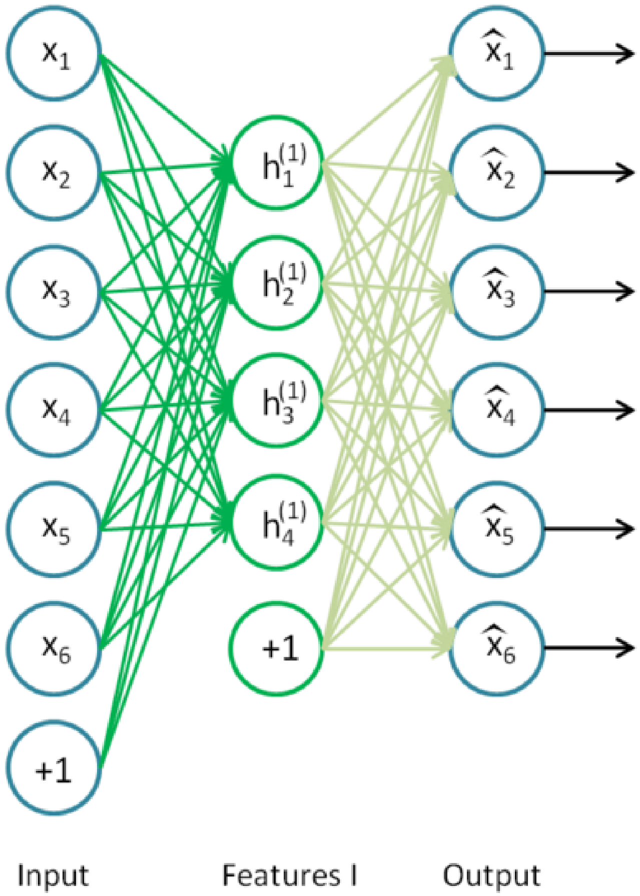


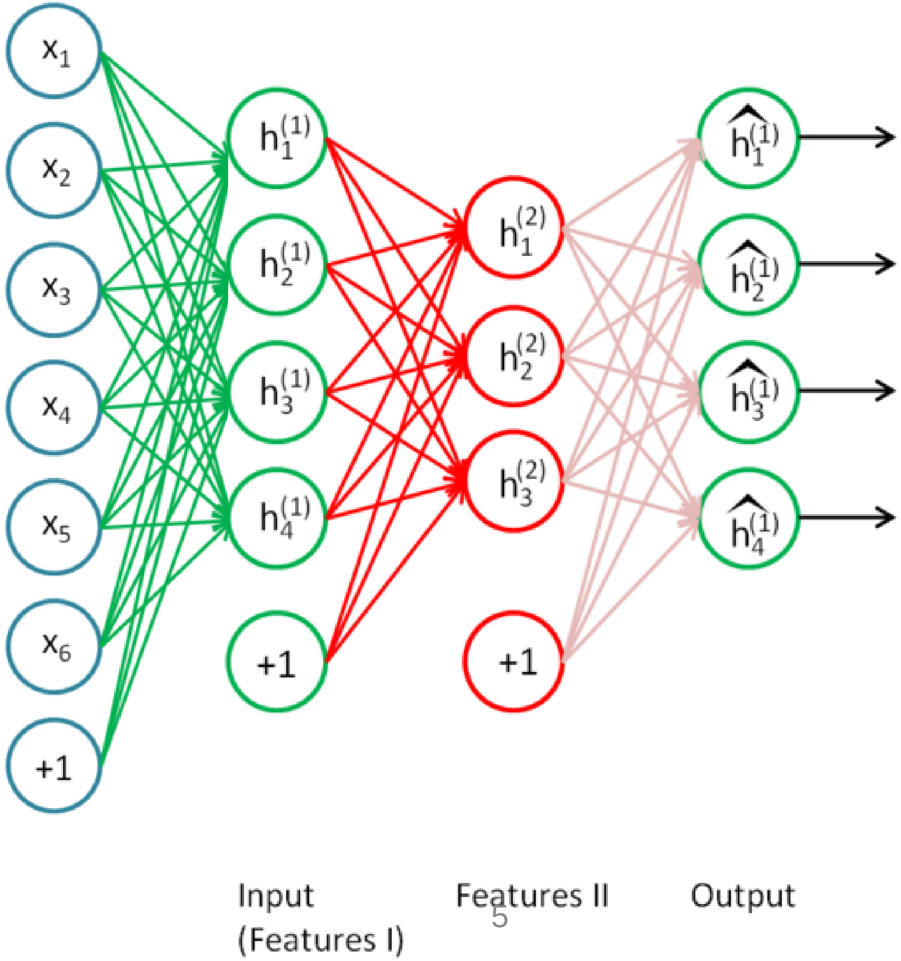
z represents the input object

- Sometimes the autoencoder with clean input only learns low level patterns instead of high level object representations
- Use noisy input instead and require the decoder to remove the noise
- To remove the noise, the decoder needs to know the object, and the encoder should learn better representations of the object
- The self-supervised pre-trained model can then be used to train on the target task

Stacked Auto-Encoders

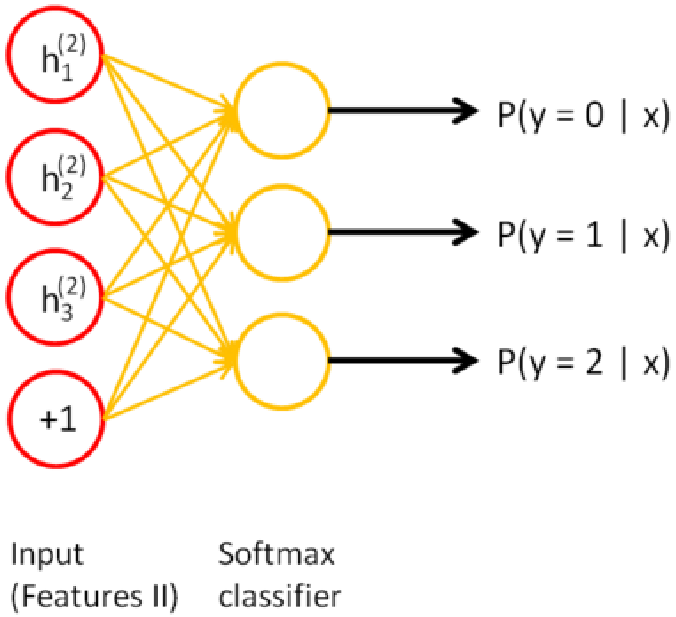
- Bengio (2007)
- Stack many (sparse) auto-encoders in succession and train them using greedy layer-wise training
- Drop the decode output layer each time





Stacked Auto-Encoders

- Do supervised training on the last layer using final features
- Then do supervised training on the entire network to fine- tune all weights



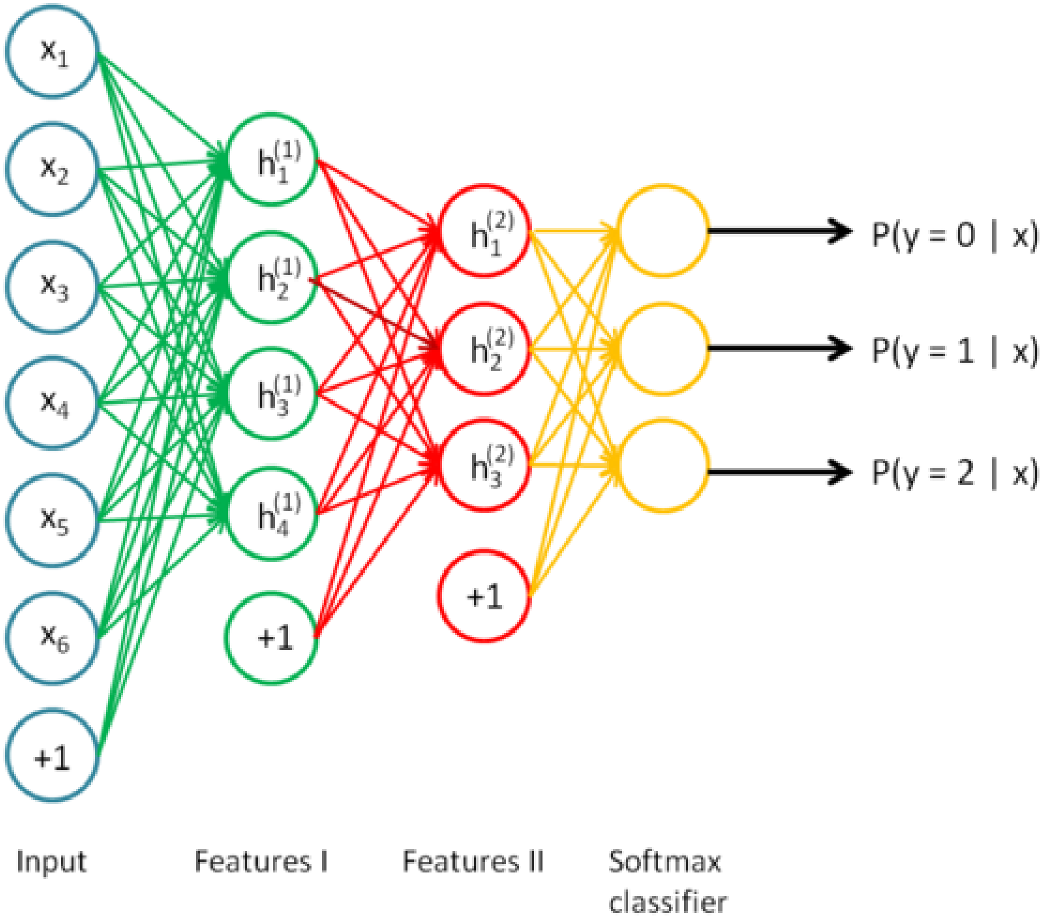
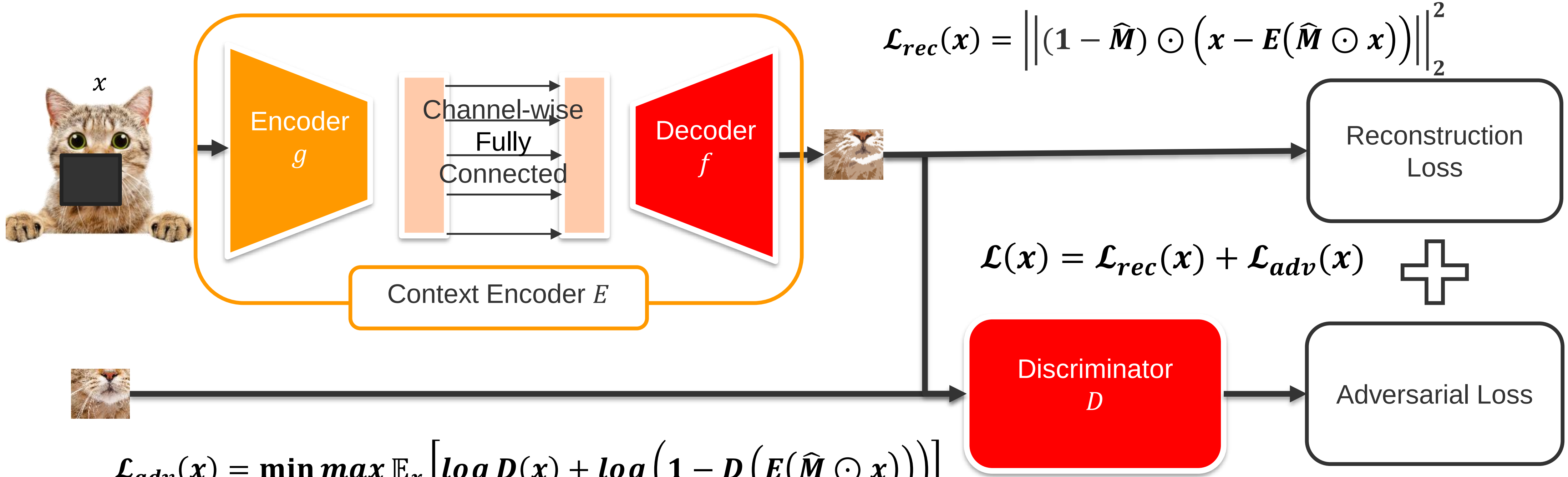


Image Inpainting

M: a bit map with masked 0, other regions 1

$$\mathcal{L}_{rec}(x) = \left\|(1 - \hat{M}) \odot \left(x - E(\hat{M} \odot x)\right)\right\|_2^2$$



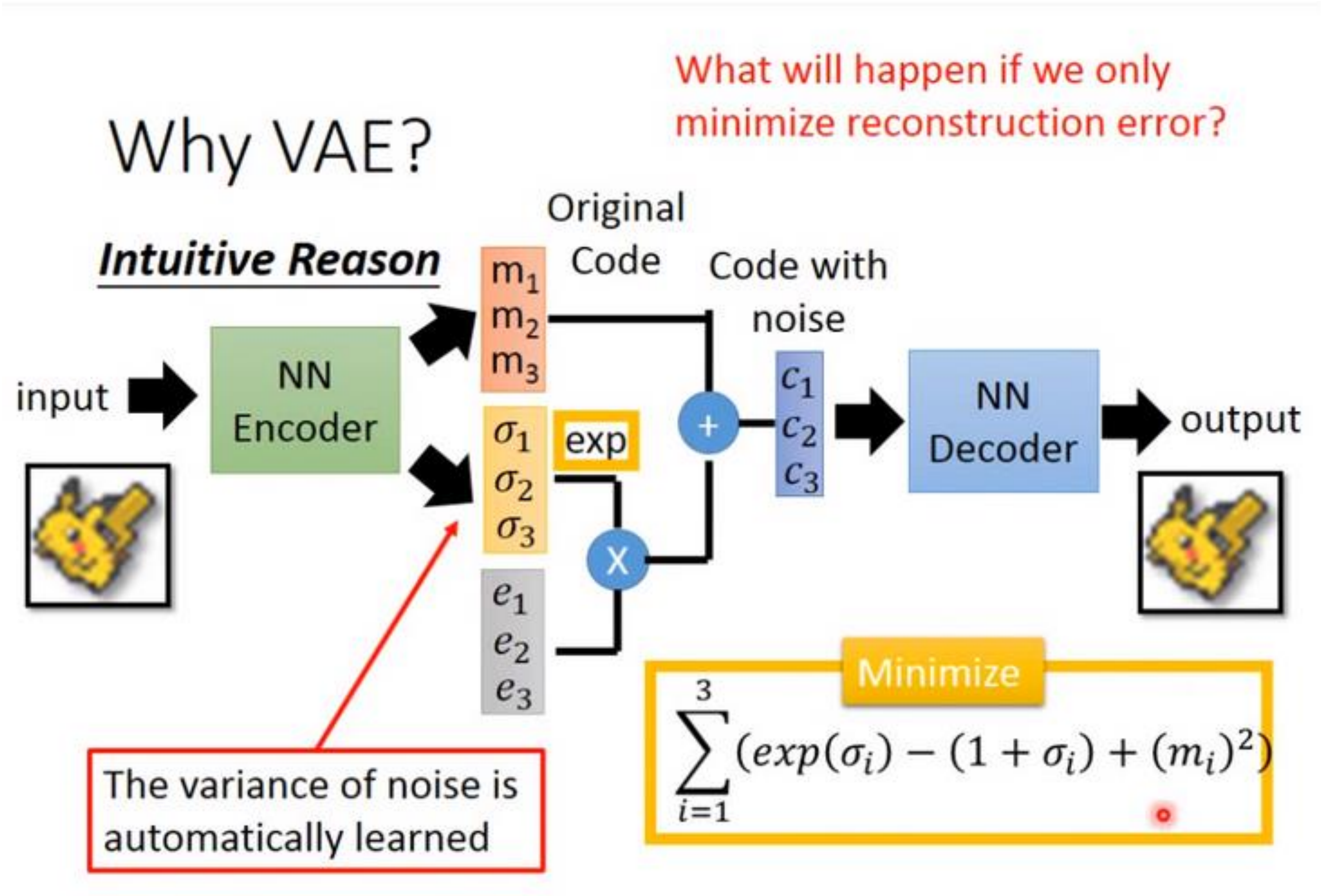
$$\mathcal{L}(x) = \mathcal{L}_{rec}(x) + \mathcal{L}_{adv}(x)$$

$$\mathcal{L}_{adv}(x) = \min_{E} \max_{D} \mathbb{E}_x \left[\log D(x) + \log\left(1 - D\left(E(\hat{M} \odot x)\right)\right)\right]$$

Pixel wise reconstruction does not give semantic reconstruction,
➔ image 較不清晰

- Part of the input image is cropped, and the decoder tries to reconstruct the missing part
- To make the reconstructed image looks more like real, a discriminator is used to identify real images and generated images, and the decoder should generate more realistic images to deceive the discriminator
- The self-supervised pre-trained model can then be used to train on the target task

VAE(Variational AutoEncoder)



Why VAE?

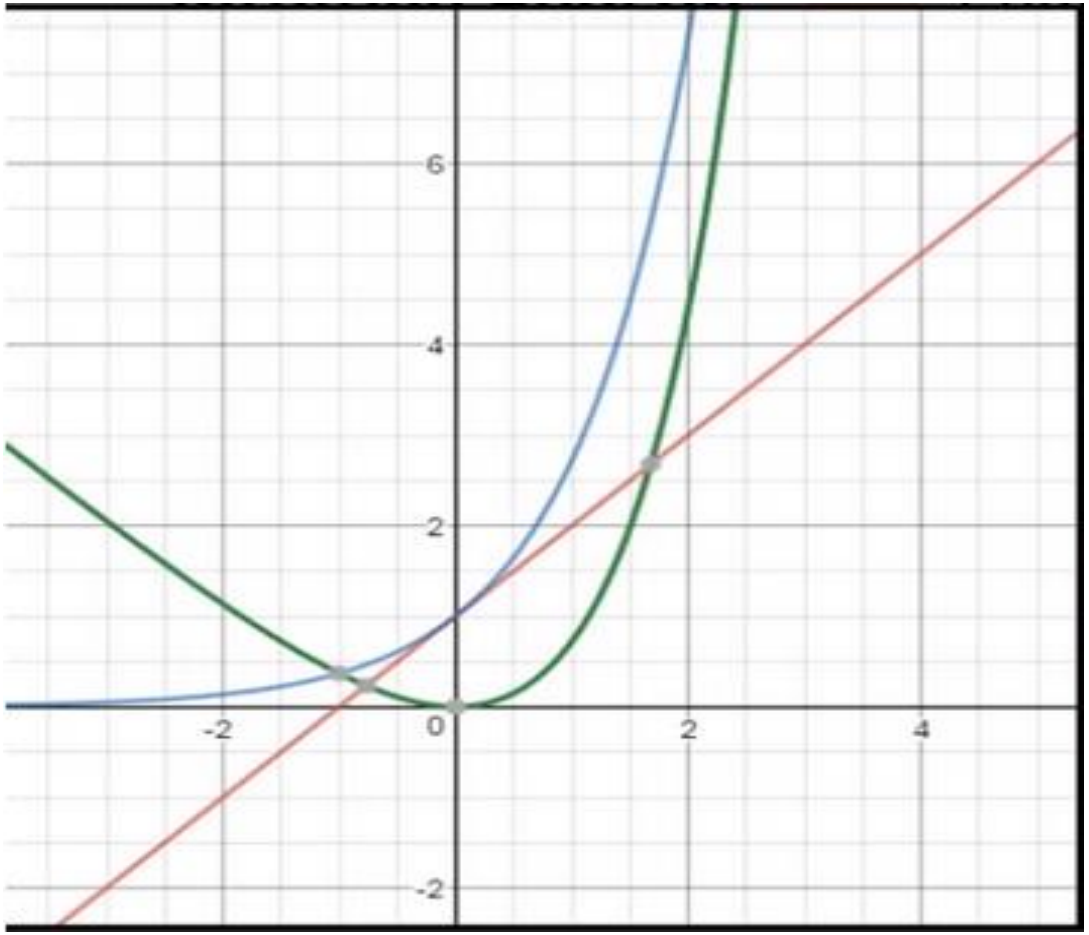
What will happen if we only minimize reconstruction error?

Intuitive Reason

The variance of noise is automatically learned

Minimize

$$\sum_{i=1}^{3}(exp(\sigma_i) - (1 + \sigma_i) + (m_i)^2)$$

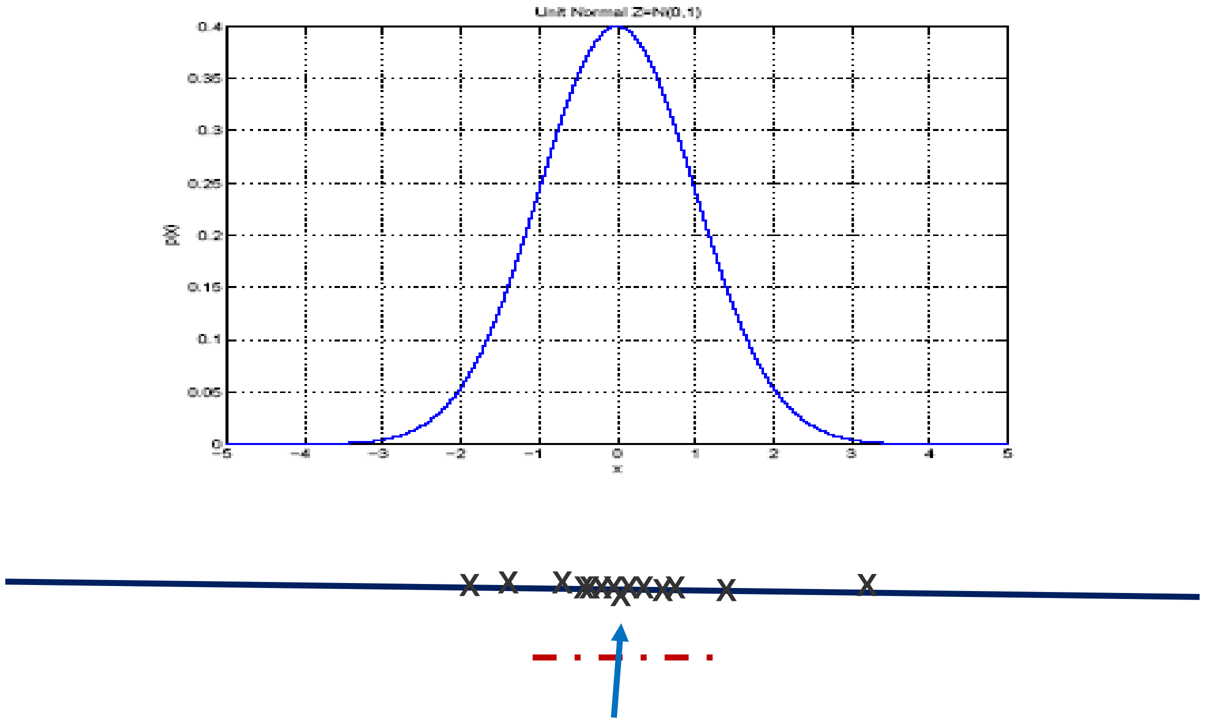


δ_i = 0, exp(δ_i)=1 has the minimum value

C_i = exp(δ_i) x e_i + m_i 依然加了 noises

Maximum likelihood estimation

When there is no reason to favor certain θ



More data points on this high prob. area

- $p(x) = \mathcal{N}(\mu_1, \sigma_1{}^2)$

Most data cannot get the high prob.
Only small amount of data have high prob.

For a correct estimation, it is expected to have the more data achieving high probability

➔Likelihood of θ given the sample $\mathcal{X}$

Maximize $p(\mathcal{X}|\vartheta) = \prod_t p(x^t|\vartheta)$

should be maximized

Bayes rule:

$$\boxed{\begin{aligned} \boldsymbol{P}(\boldsymbol{\theta}\,|\,\boldsymbol{A}) &= \frac{\boldsymbol{P}(\boldsymbol{\theta}\cap\boldsymbol{A})}{\boldsymbol{P}(\boldsymbol{A})} \\ &= \frac{\boldsymbol{P}(\boldsymbol{A}\,|\,\boldsymbol{\theta})\boldsymbol{P}(\boldsymbol{\theta})}{\boldsymbol{P}(\boldsymbol{A})} \end{aligned}}$$

- To model X with parametrized distribution P_θ
- Let Z represent a latent encoding of X
- P_θ(x,z) represents the joint distribution under P_θ of the observable data and its latent space z, where

$$p_\theta(\mathbf{x}) = \int_{\mathbf{z}} p_\theta(\mathbf{x} \mid \mathbf{z}) p_\theta(\mathbf{z})\, d\mathbf{z}$$

Define the set of relationships between the input data and its latent representation:

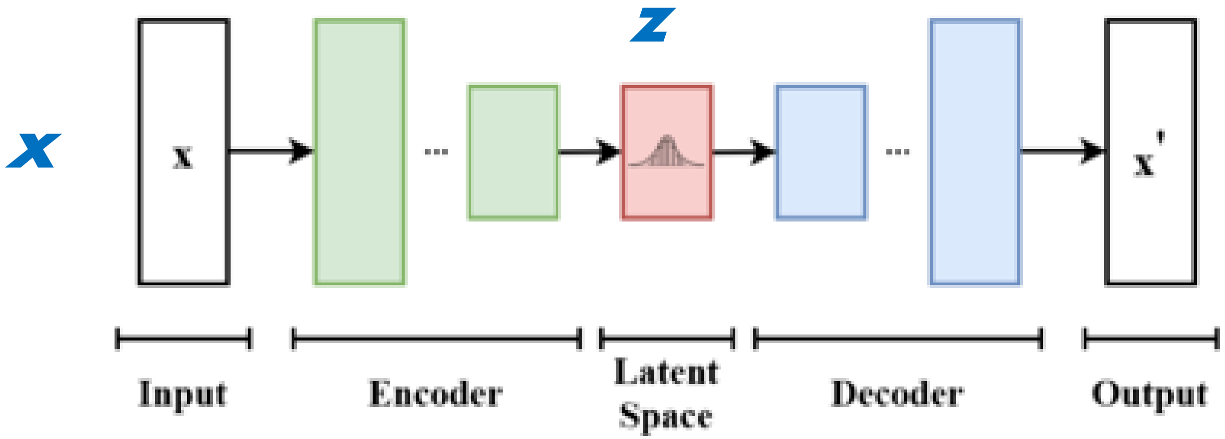
- Prior $p_\theta(\mathbf{z})$
- Likelihood $p_\theta(\mathbf{x} \mid \mathbf{z})$
- Posterior $p_\theta(\mathbf{z} \mid \mathbf{x})$

找個模型來實作 coder

$q_\emptyset(z|x) \approx p_\theta(z|x)$

但我們擁有的只有所有的 X

Or, P(x)



Z is the code,
in the latent space

$$\mathrm{D}_{\mathrm{KL}}(q_{\emptyset}(z|x)\|p_{\theta}(z|x)) = \int (q_{\emptyset}(z|x)\log\frac{q_{\emptyset}(z|x)}{p_{\theta}(z|x)}dz \quad = \int (q_{\emptyset}(z|x)\log\frac{q_{\emptyset}(z|x)p_{\theta}(x)}{p_{\theta}(z,x)}dz$$

$$= \int (q_{\emptyset}(z|x)\left(\log(p_{\theta}(x)) + \log\frac{q_{\emptyset}(z|x)}{p_{\theta}(z,x)}\right)dz$$

$$= \log(p_{\theta}(x)) + \int q_{\emptyset}(z|x)\log\frac{q_{\emptyset}(Z|X)}{p_{\theta}(z,x)}dz$$

$$\Rightarrow \quad \log(p_{\theta}(x)) = -\int q_{\emptyset}(z|x)\log\frac{q_{\emptyset}(Z|X)}{p_{\theta}(z,x)}dz + \mathrm{D}_{\mathrm{KL}}(q_{\emptyset}(z|x)\|p_{\theta}(z|x))$$

Maximize ： $\underbrace{\log(p_\theta(x))}_{\text{Max}} - \underbrace{D_{KL}(q_\emptyset(z|x)\|p_\theta(z|x))}_{\text{Min}}$

$$\boxed{= \underbrace{E_{Z\sim q_\emptyset(z|x)}\left(\log\left(p_\theta(x|z)\right)\right)}_{\text{Max}} - \underbrace{D_{KL}(q_\emptyset(z|x)\|p_\theta(z))}_{\text{Min}}}$$

lower bound

Minimize ：

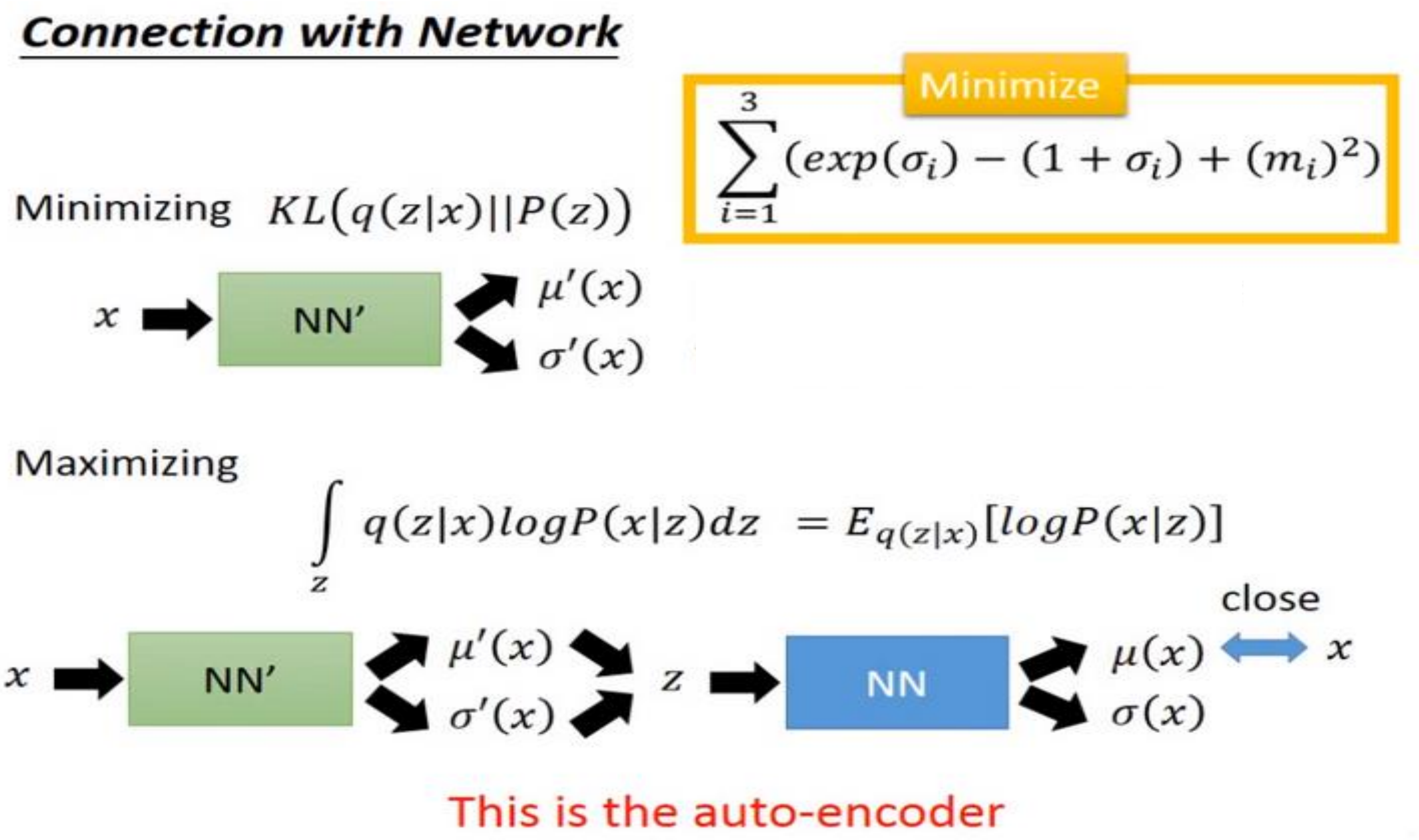
$$\begin{aligned}\underline{L_{\theta,\emptyset}} &= -\log(p_\theta(x)) + D_{KL}(q_\emptyset(z|x)\|p_\theta(z|x)) \\ &= -E_{Z\sim q_\emptyset(z|x)}\left(\log\left(p_\theta(z|x)\right)\right) + D_{KL}(q_\emptyset(z|x)\|p_\theta(z))\end{aligned}$$

Evidence lower bound (ELBO) loss function

$$\theta^*, \emptyset^* = \underset{\theta,\emptyset}{\text{argmin}}\ L_{\theta,\emptyset}$$

$$\boxed{\begin{array}{c} -L_{\theta,\emptyset} = \log\left((p_\theta(x))\right) - D_{KL}(q_\emptyset(z|x)\|p_\theta(z|x)) \leq \log(p_\theta(x)) \\ KL\text{必定} \geq 0 \qquad \text{=0 when } q_\emptyset(z|x) = p_\theta(z|x) \end{array}}$$

Maximizing Likelihood



Connection with Network

Minimizing $KL\big(q(z|x)||P(z)\big)$

Minimize

$$\sum_{i=1}^{3}\left(exp(\sigma_i) - (1 + \sigma_i) + (m_i)^2\right)$$

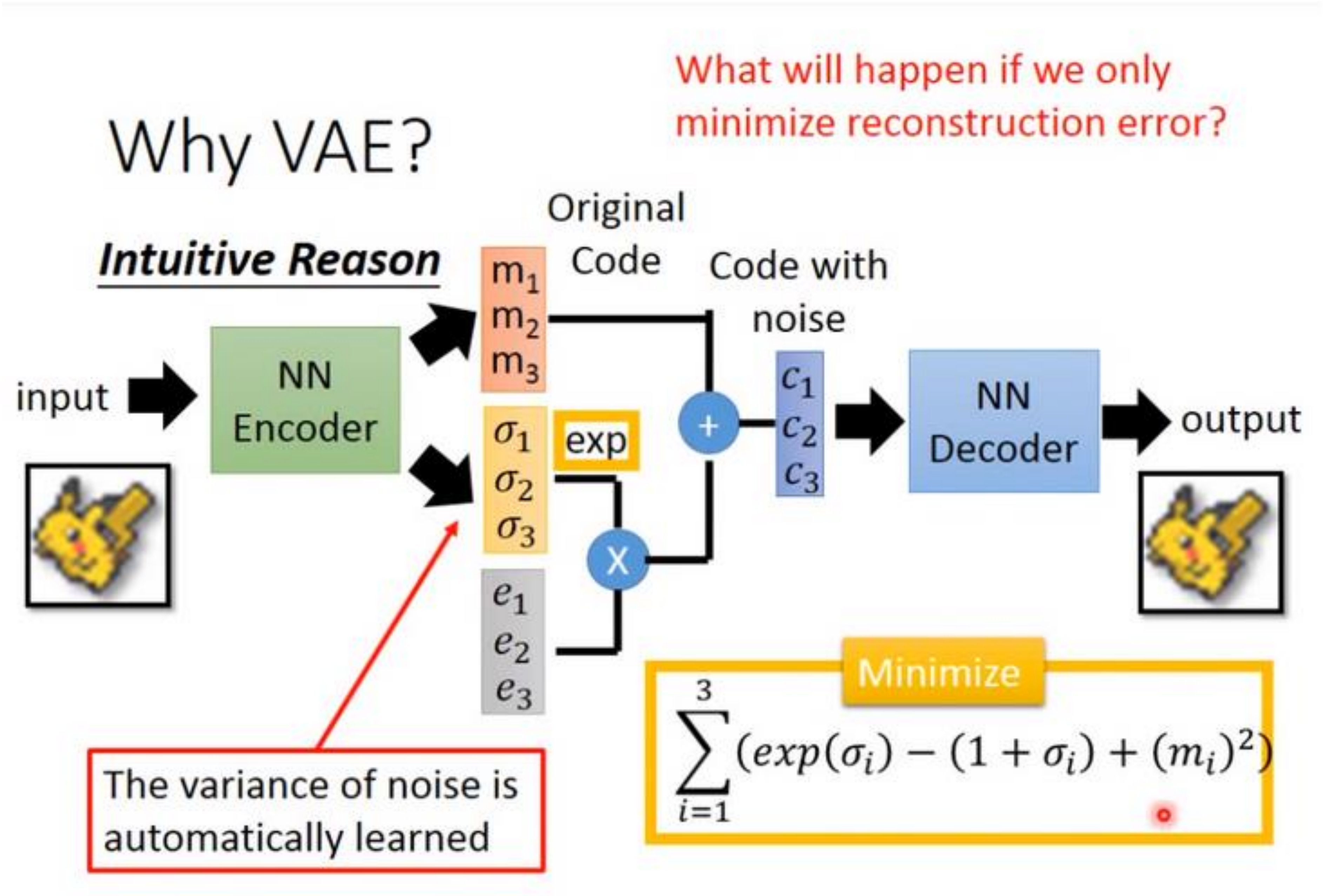
x
NN’
$\mu'(x)$
$\sigma'(x)$

Maximizing

$$\int_z q(z|x)logP(x|z)dz \;= E_{q(z|x)}[logP(x|z)]$$

x
NN’
$\mu'(x)$
$\sigma'(x)$
z
NN
$\mu(x)$
$\sigma(x)$
close
x

This is the auto-encoder



Why VAE?

What will happen if we only minimize reconstruction error?

Intuitive Reason

The variance of noise is automatically learned

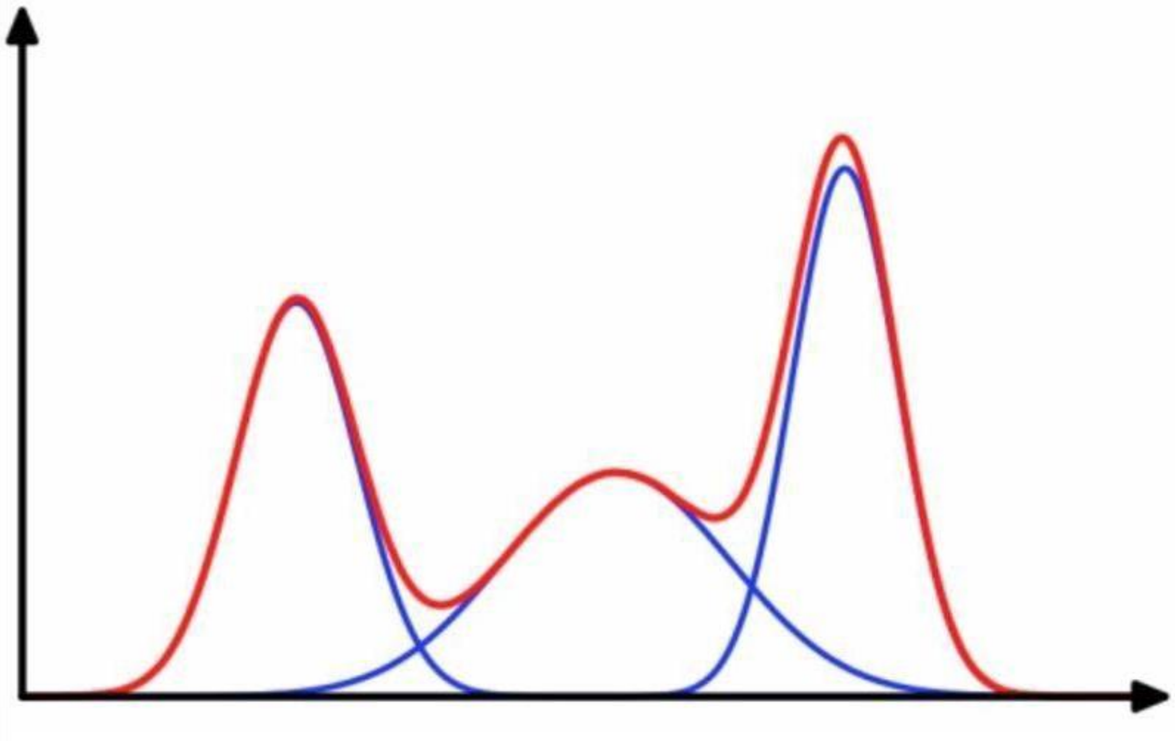
Minimize

$$\sum_{i=1}^{3}(exp(\sigma_i) - (1 + \sigma_i) + (m_i)^2)$$

Gaussian mixture Model

- GMM(Gaussian mixture Model)
- Considers each cluster as a different Gaussian distribution
- Mixture different distribution (blue line)

  into new distribution(red line)

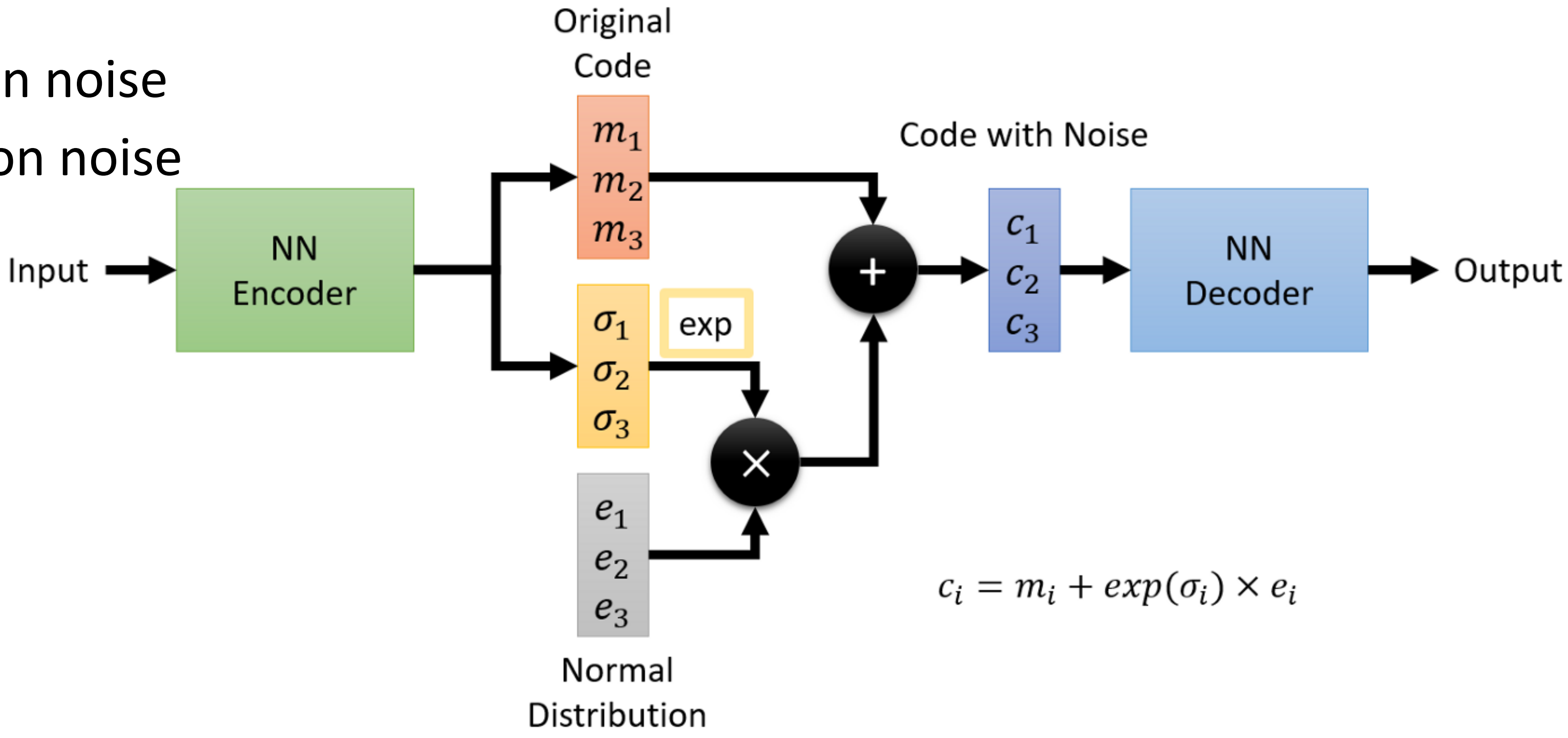


VAE structure

利用 Normal distribution抽樣

σ :control weight of
normal distribution noise

e: normal distribution noise



$$c_i = m_i + exp(\sigma_i) \times e_i$$

VAE structure

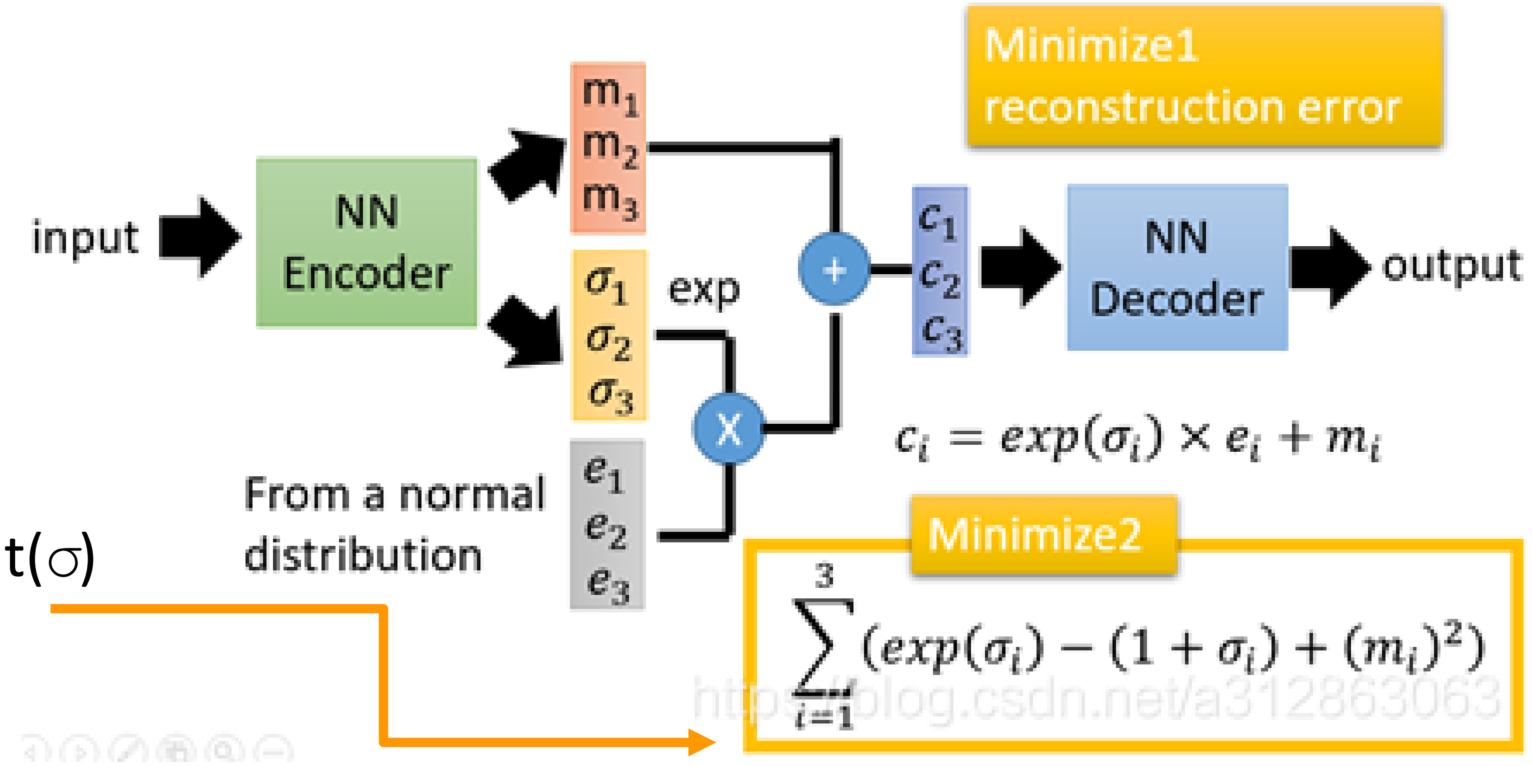
σ:controlweightof
normal distribution noise

e: normal distribution noise

Two loss for training

1.Reconstruction loss

2.Bounding of “noise” weight(σ)



Appendix:

$$\mathrm{D}_{\mathrm{KL}}(q_{\emptyset}(z|x)\|p_{\theta}(z|x)) = \int (q_{\emptyset}(z|x)\log\frac{q_{\emptyset}(z|x)}{p_{\theta}(z|x)}dz$$

$$= \log(p_{\theta}(x))+\int q_{\emptyset}(z|x)\log\frac{q_{\emptyset}(Z|X)}{p_{\theta}(z,x)}dz$$

$$= \log(p_{\theta}(x))+\int q_{\emptyset}(z|x)\log\frac{q_{\emptyset}(Z|X)}{\underline{p_{\theta}(X|Z)}p_{\theta}(z)}dz$$

$$= \log(p_{\theta}(x))+E_{Z\sim q_{\emptyset}(Z|X)}\left(\log\frac{q_{\emptyset}(Z|X)}{p_{\theta}(z)} - \underline{\log(p_{\theta}(x|z))}\right)$$

$$= \log(p_{\theta}(x))+E_{Z\sim q_{\emptyset}(Z|X)}\underline{\Bigg(\log\frac{q_{\emptyset}(Z|X)}{p_{\theta}(z)}} - \log(p_{\theta})(x|z)\Bigg)$$

$$= \log(p_{\theta}(x))+\underline{D_{KL}(q_{\emptyset}(z|x)\|p_{\theta}(z))} - E_{Z\sim q_{\emptyset}(Z|X)}\big(\log(p_{\theta}(x|z))\big)$$