$$P(y|x) = \frac{P(x|y)P(y)}{P(x)} \quad \Big| \quad P(x|y) = \frac{P(y|x)P(x)}{P(y)}$$

$$P(y|\underline{x})$$

Generative Models

(Recap of Autoencoder, Variational Autoencoder)

$P(y|x) \propto P(x|y)\, P(y)$

$P(x|y) \propto \underbrace{\boxed{P(y|x)\, P(x)}}_{P(x,y)}$

Tirtharaj Dash

Dept. of CS & IS and APPCAIR
BITS Pilani, Goa Campus

December 02, 2021

Discriminative Model

$P(y|\underline{x}) \leftarrow$

Generative Model ?

$P(\underline{x}, y)$

$P(\underline{x}) \quad P(x_1, x_2, x_3, \ldots, x_d)$

different kinds of $\underline{x}$

① $\underline{x}$: Tabular structure — every row in the table rep. an instance

MLP

② $\underline{x}$: visual data (spatial)

CNN

③ $\underline{x}$: Temporal data

RNN, Tran. (sequences)

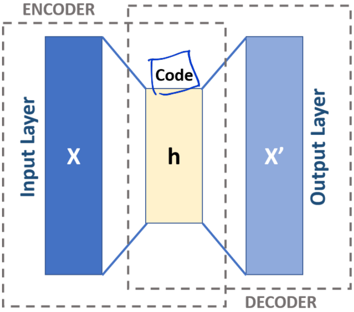
④ $\underline{x}$: Graph-structured (relational) data

GNN

Autoencoder – Discriminative Model

reconstruct

- AE is a neural network that learns to "copy" its input to its output.
- It has an internal (hidden) layer that describes a **code** used to represent the input.



Autoencoder – Discriminative Model

- It is constituted by two main parts: an encoder that maps the input into the code, and a decoder that maps the code to a reconstruction of the input.
- However, simply copying input to output would just *duplicate* the singal (rather than generalising). Why?
- Instead, AE reconstructs the input approximately, preserving the most relevant aspects of the data (we can call this: some important latent aspects).

Autoencoder – Discriminative Model

- Let $\mathbf{x}$ be an input example. The encoder and decoder do the following:

$$Enc : \mathcal{X} \mapsto \mathcal{H}$$

$$Dec : \mathcal{H} \mapsto \mathcal{X}$$

- Where, *Enc* and *Dec* are the functions obtained by minimising a recostruction loss.

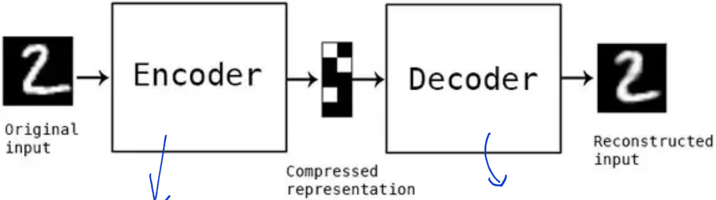
$\mathcal{L}(x, \hat{x})$

$$Enc, Dec = \arg \min_{Enc, Dec} ||\mathbf{x} - (Dec \circ Enc)(\mathbf{x})||^2$$

- In the simplest case, both encoder and decoder are single layered. That is:

$$\mathbf{h} = \sigma(\mathbf{wx} + b)$$

- **h** is referred to as code or latent variable or latent representation.



CNN

CPCPCP...

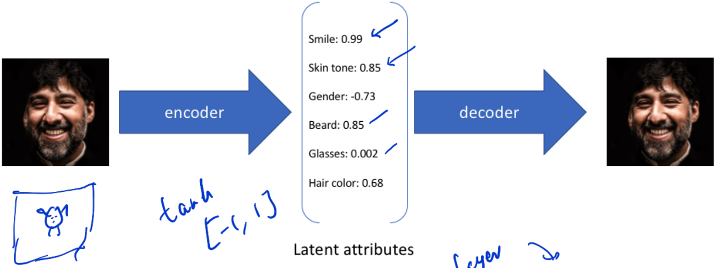
or CLCC...

Deconv.

CCCC...

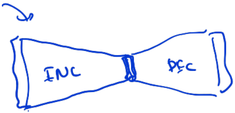
Autoencoder – Discriminative Model

- Each hidden dimension represents some latent feature learned about the input.
- For example (the features mentioned are hypothetical for demonstration purpose only):



(Fig source: Jeremy Jordan's blog)

Bottleneck Layer
or Code Layer



Autoencoder – Discriminative Model

- This is a discriminative model.
- AE can be used for:
  - compressing data
  - greedy layerwise pre-training
- AE cannot used for: <u>generating new data</u>.
- For generating new data, the model needs to learn a joint distribution of some kind $p(\mathbf{x})$ or $p(\mathbf{x}, \mathbf{h})$.
- Or, a model can be considered as "generative" when the input latent variable has probability distribution associated with it – a kind of autoencoder that does this is 'Variational AE'.

$p(\underline{h} \mid \underline{x})$ image, code

$p(\underline{h} \mid \underline{x})$

$\underline{x}$	$\underline{h}$
im1	[0.5, 0.6]
im2	[0.2, 0.8]
im3	[0.4, 0.9]
im4	[0.7, -0.1]
im5	[0.1, 0.9]
?	← [0.35, 0.75]

$p(\underline{x}, \underline{h})$

$p(\underline{x})$

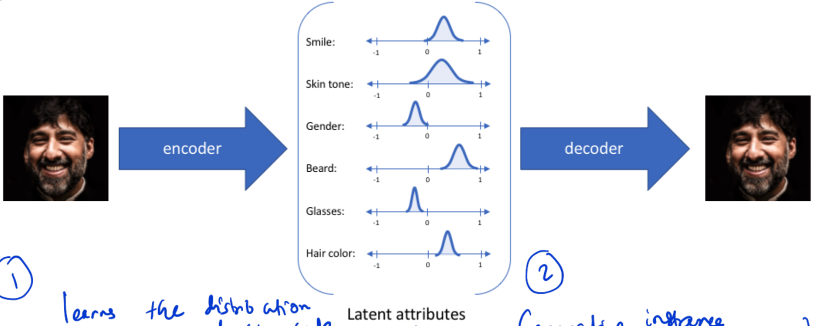
? ←

AE

Given $\underline{X}$, learns ~~a point estimate~~ $\underline{h}$ ← distribution.

"Variational" AE : Generative AE

- VAE imposes a specific probabilistic structure on the hidden units.



1 learns the distribution of the code

$p(x, w)$

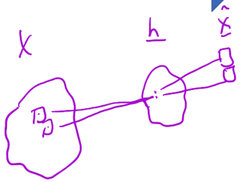
2 Generating instance

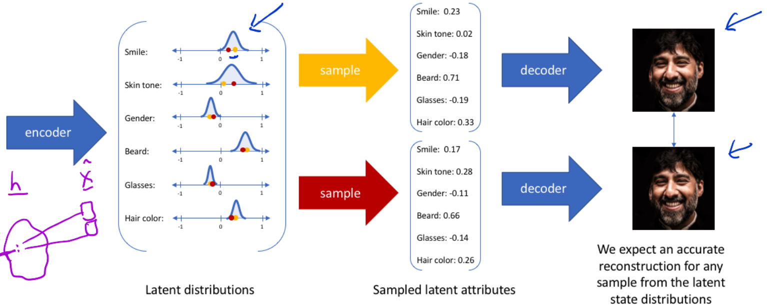
(Fig source: Jeremy Jordan's blog)

- The AE network is sometimes called 'recognition model' whereas the decoder network is sometimes referred to as the 'generative model'.

VAE

- The encoder model outputs a range of possible values (a statistical distro) from which, we can randomly sample and input to the decoder to re-construct the input. This enforces a continuous and smooth latent space representation.
- It is expected that values that are close enough in the latent space would result in similar reconstructions.





VAE

(Fig source: Jeremy Jordan's blog)

- Suppose there exists some hidden variable $\mathbf{z}$ which generates an observation $\mathbf{x}$. As a Bayesian network, it looks like:

- The difficulty is that we know nothing of $\mathbf{z}$, we can only see $\mathbf{x}$. But, we can infer some characteristics of $\mathbf{z}$:

likelihood

$$p(\mathbf{z}|\mathbf{x}) = \frac{p(\mathbf{x}|\mathbf{z})p(\mathbf{z})}{p(\mathbf{x})} \quad : P(\underline{x}, \underline{z})$$

prior

marginal

$X \to \boxed{z} \to$

unknown.

$p(\underline{z}|\underline{x}) \propto p(\underline{x}|\underline{z})\, P(\underline{z})$

$= \int_z P(x|z)P(z)\, dz$

VAE

Exact estimate for $p(z|x)$ is not possible.

↓

approximate if

oblem Computing the denominator $p(\mathbf{x})$ is hard:

$$p(\mathbf{x}) = \int \boxed{p(\mathbf{x}|\mathbf{z})} p(\mathbf{z}) d\mathbf{z}$$

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

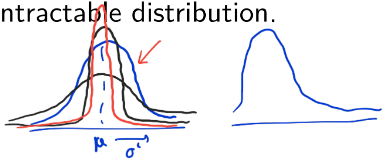
lution Variational inference

- Let's approximate $p(\mathbf{z}|\mathbf{x})$ by another distribution $q(\mathbf{z}|\mathbf{x})$ such that it has a tractable distribution.
- If we can define the parameters of $\boxed{q(\mathbf{z}|\mathbf{x})}$ such that it is very similar to $p(\mathbf{z}|\mathbf{x})$, we can use it to perform approximate inference of the intractable distribution.

$p \xrightarrow{①} q$

$\downarrow ②$

params ?



VAE

$p \xrightarrow[\text{distance}]{} q$

$D_{KL}(p \| q) = \sum_{x \in X} p_i(x) \log \frac{p_i(x)}{q_i(x)}$

- If we minimise KL divergence between $q(\mathbf{z}|\mathbf{x})$ and $p(\mathbf{z}|\mathbf{x})$, then these two distros will become similar to each other.

$\sum q(z|x) \log \frac{q(z|x)}{p(z|x)}$

$$\min KL(q(\mathbf{z}|\mathbf{x})||p(\mathbf{z}|\mathbf{x}))$$

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

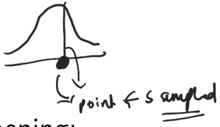
Which in turn can be solved by maximising the following (not showing the proof here):

$$\boxed{E_{q(\mathbf{z}|\mathbf{x})} \log p(\mathbf{x}|\mathbf{z}) - KL(q(\mathbf{z}|\mathbf{x})||p(\mathbf{z}))}$$

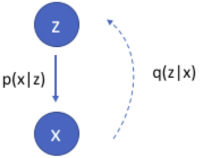
The first term represents reconstruction likelihood. The second term enforces that q distro is similar to the true distro $p(\mathbf{z})$.

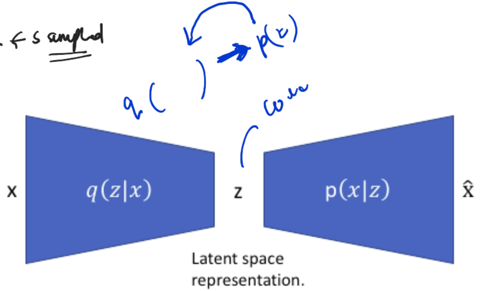
$\underline{x} \longrightarrow \boxed{z} \longrightarrow \underline{\hat{x}}$

VAE



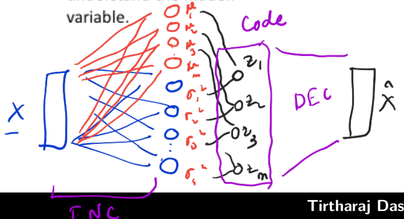
- This is what is happening:





We'd like to use our observations to understand the hidden variable.

$N($ $)$



$p(z)$

$\downarrow$

$q(z|x)$ — Parameters

$q : \mathcal{N}(\mu, \sigma^2)$