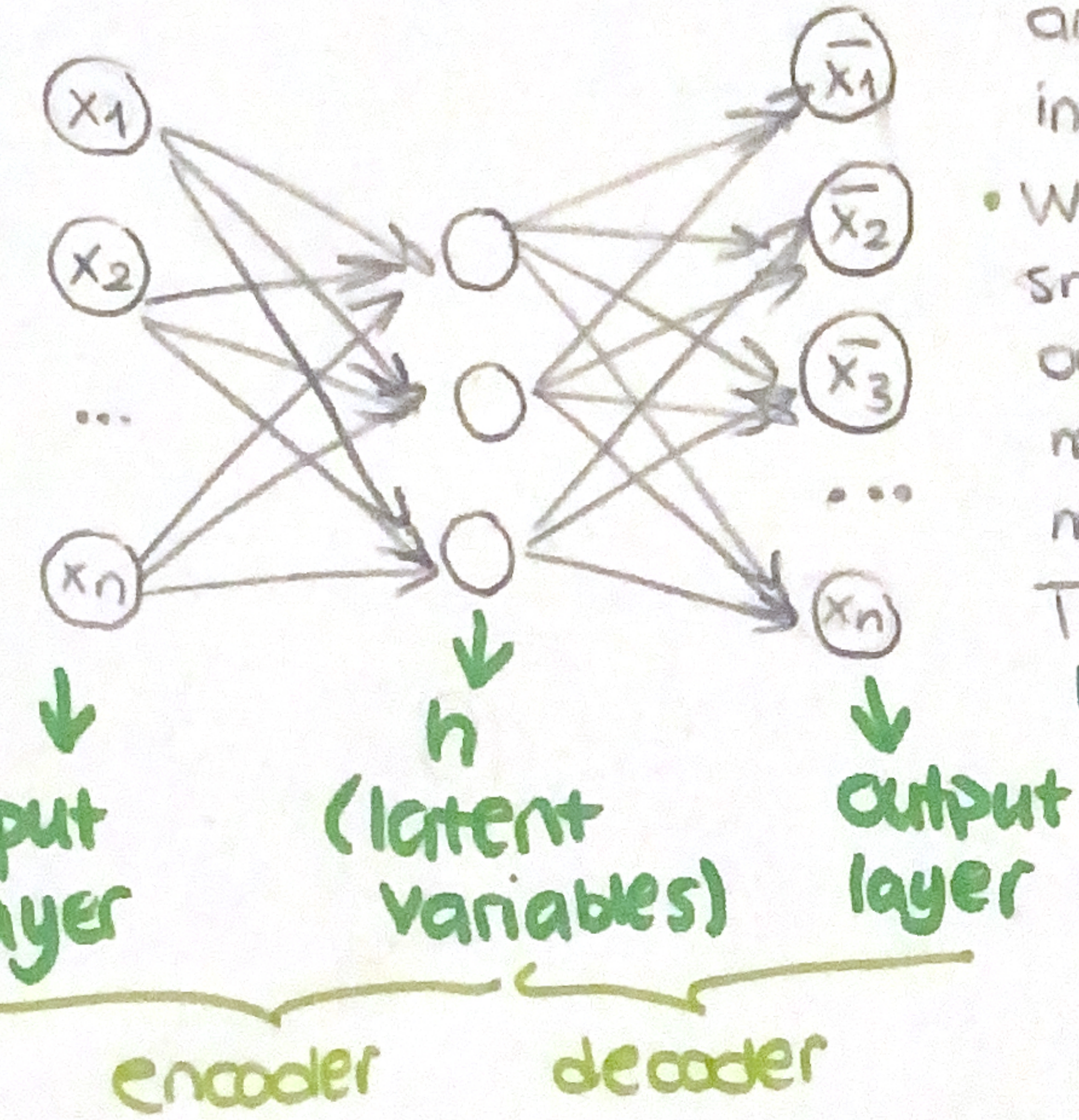
Autoencoders



Idea: Autoencoders are trained to reconstruct input.

- When hidden layer is smaller than input & output layers, the model only learns the most salient features. This is called undercomplete autoencoder.
- Autoencoders learn unsupervised.
- $h_{w,b}(x) \approx x$ → identity function (were trying to learn this)

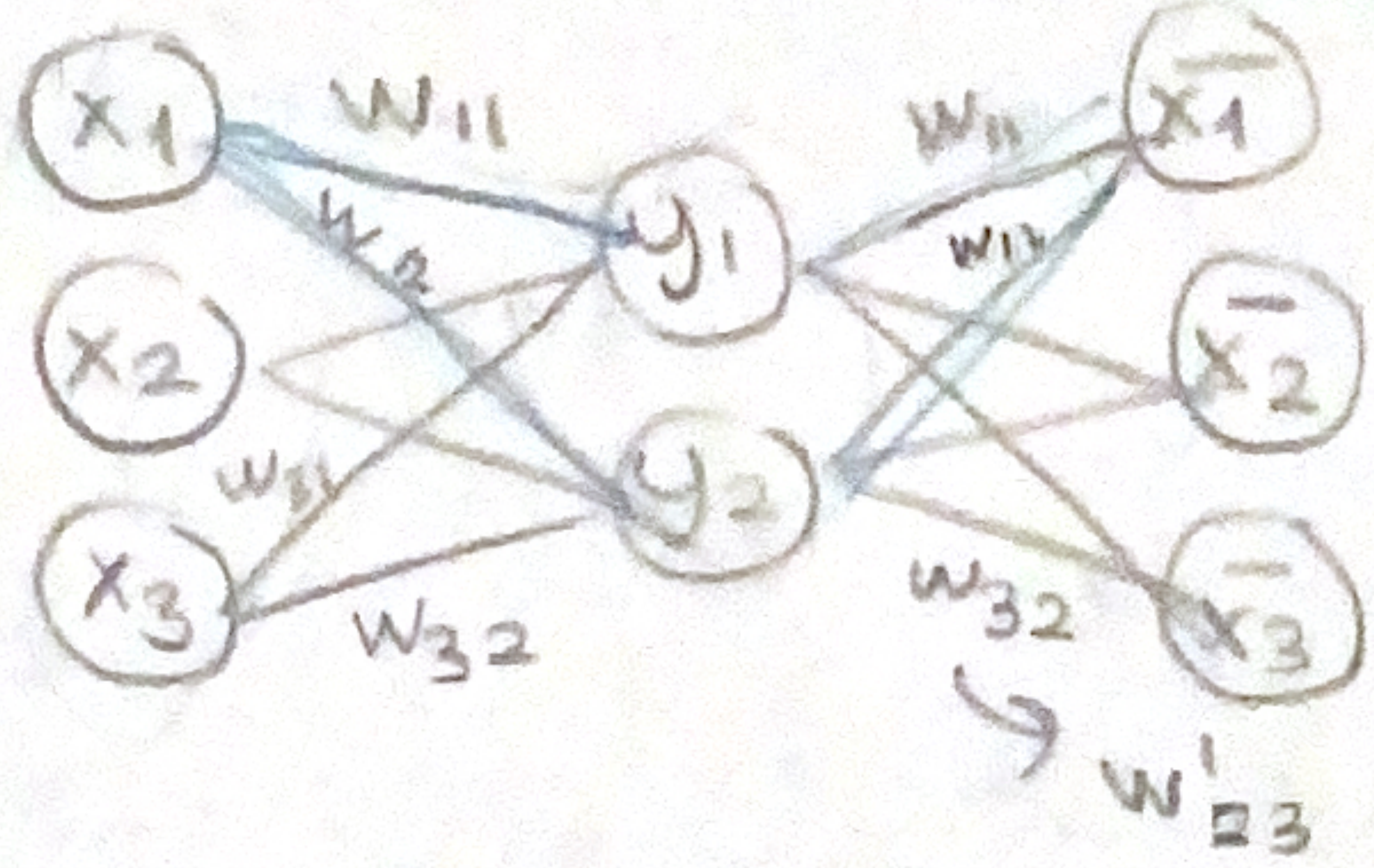
LOSS → $(\bar{x}_i - x_i)$

$x_i \xrightarrow[b]{w} h \to \bar{x}_i$ → output $= \bar{x} = S(w'y+b')$ ↓ ↓ these are weights and biases of decoder

↳ output $= y = S(wx+b)$ ↓ ↓ weights and biases of encoder

L (x - J(w,x,)

$W' = W^T$ ← one is transpose of another!



$$W = \begin{array}{c} \\ \\ \end{array}\begin{array}{ccc} x_1 & x_2 & x_3 \\ \end{array}$$

$$W = \left[\begin{array}{ccc} w_{11} & w_{21} & w_{31} \\ w_{12} & w_{22} & w_{32} \end{array}\right]$$

$$W' = W^T \to W' = \begin{bmatrix} w_{11} & w_{12} \\ w_{21} & w_{22} \\ w_{31} & w_{32} \end{bmatrix}$$

$$L(x, \bar{x}) = \|x - \bar{x}\|^2$$