MINST dataset.

787	28
28	

m training images.

28X28.

$$X = \begin{bmatrix} - & x^{(1)} & - \\ - & x^{(2)} & - \\ & \vdots & \\ - & x^{(n)} & - \end{bmatrix}^T = \begin{bmatrix} | & | & & | \\ x^{(1)} & x^{(2)} & \cdots & x^{(n)} \\ | & | & & | \end{bmatrix}$$

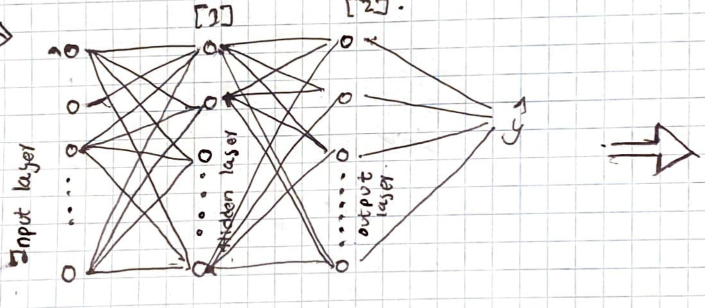
784

28x28

$\Downarrow$

$$\underbrace{0, 1, 2 \cdots\cdots 9}_{10 \text{ classes.}}$$

Forward propagation



$A^{[0]} = X \quad (784 \times m) \cdots (i)$

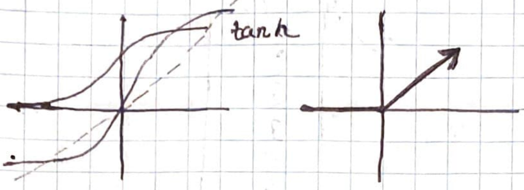
$Z^{[1]} = W^{[1]} A^{[0]} + b^{[1]} : (ii)$

$\begin{matrix} 10 \times m & 10 \times 784 & 784 \times m & 10 \times 1 \end{matrix}$

$A^{[1]} = g(Z^{[1]}) = ReLU(Z^{[1]}) \cdots (iii)$

2nd layer is linear combination of input layers

rectified linear unit.



ReLU(x)

$$\rightarrow \begin{cases} x \text{ if } x > 0 \\ 0 \text{ if } x \leq 0 \end{cases}$$

$$\underset{10 \times m}{Z^{[2]}} = \underset{10 \times 10}{W^{[2]}} \; \underset{10 \times m}{A^{[1]}} + \underset{10 \times 1}{b^{[2]}} \quad \cdots (\text{iv}) \quad \Rightarrow 10 \times m.$$

$$A^{[2]} = \text{softmax}\left(Z^{[2]}\right) \cdots (\text{v}).$$

Each of the 10 nodes correspond to each of the 10 digits that could be recognized. need a probability

Output layer

softmax activation.

$$\begin{bmatrix} 1.3 \\ 5.1 \\ 2.2 \\ 0.7 \\ 1.1 \end{bmatrix} \longrightarrow \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}} \longrightarrow \begin{bmatrix} 0.2 \\ 0.90 \\ 0.05 \\ 0.01 \\ 0.02 \end{bmatrix}$$

(vi).

{Back word propagation}

on (iv)

one hot encode

$y = 4 \Rightarrow \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ \vdots \end{bmatrix}$

$$dZ^{[2]} = A^{[2]} - Y$$

10xm 10xm 10xm.

$$dW^{[2]} = \frac{1}{m} dZ^{[2]} A^{[1]T}$$

10x10 10xm mx10

$$db^{[2]} = \frac{1}{m} \sum dZ^{[2]}$$

10x1 10x1

$$dZ^{[1]} = W^{[2]T} dZ^{[2]} .* g'(Z^{[1]})$$

10xm 10x10 10xm 10xm

$$dW^{[1]} = \frac{1}{m} dZ^{[1]} X^{T}$$

10x784 10xm mx784

$$db^{[2]} = \frac{1}{m} \sum dZ^{[1]}$$

10x1 10x1

α = learning rate

$$W^{[1]} = W^{[1]} - \alpha dW^{[2]}$$

$$b^{[1]} = b^{[1]} - \alpha db^{[1]}$$

$$W^{[2]} = W^{[2]} - \alpha dW^{[2]}$$

$$b^{[2]} = b^{[2]} - \alpha db^{[2]}$$

Forward propagation Backword propagation