Practicals

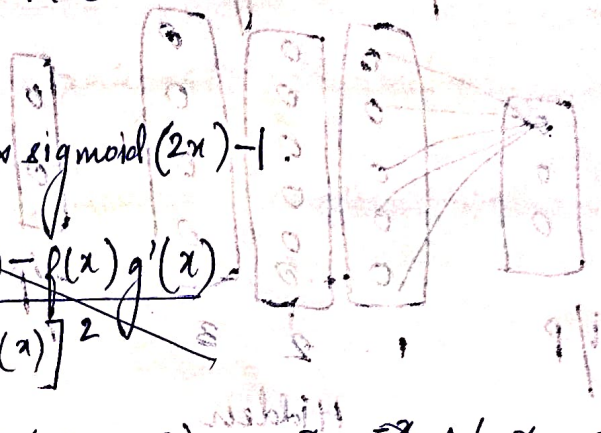
Lab 1

derivative

$$\tanh = \frac{e^x - e^{-x}}{e^x + e^{-x}} \Rightarrow \frac{2}{1+e^{-2x}} - 1$$

(*)

$$\tanh(x) = 2 \times \text{sigmoid}(2x) - 1$$



$$\frac{d}{dx}\left(\frac{f(x)}{g(x)}\right) = \frac{f'(x)g(x) - f(x)g'(x)}{[g(x)]^2}$$

$$= \frac{e^x + e^{-x} \frac{d}{dx}(e^x - e^{-x}) - e^x - e^{-x} \frac{d}{dx}(e^x + e^{-x})}{(e^x + e^{-x})^2}$$

$$=$$

$$\text{Relu}(x) = x^+ = \max(0, x) = \frac{x + |x|}{2} = \begin{cases} x & \text{if } x > 0, \\ 0 & x \leq 0 \end{cases}$$

derivative.

(*) $$\frac{e^x + e^{-x} \frac{d}{dx}(e^x - e^{-x}) - e^x - e^{-x} \frac{d}{dx}(e^x + e^{-x})}{(e^x + e^{-x})^2}$$

$$\Rightarrow \frac{(e^x + e^{-x})^2 - (e^x - e^{-x})^2}{(e^x + e^{-x})^2}$$

$$\Rightarrow 1 - \frac{(e^x - e^{-x})^2}{(e^x + e^{-x})^2}$$

$$\Rightarrow 1 - \left[\frac{e^x - e^{-x}}{e^x - e^{-x}}\right]^2$$

Leaky Relu

$$(x) = \begin{cases} x & \text{if } x > 0 \\ ax & \text{if } x \leq 0 \end{cases}$$

$$\text{Leaky Relu}(x) = \frac{(1+\alpha)x + (1-\alpha)x}{2}$$

Softmax $\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \Rightarrow \begin{bmatrix} s_1 \\ s_2 \\ s_3 \end{bmatrix}$

$x = \begin{bmatrix} 2.0 \\ 1.0 \\ 0.1 \end{bmatrix} \Rightarrow s = \begin{bmatrix} s_1 \\ s_2 \\ s_3 \end{bmatrix}$ $\quad s_i = \frac{e^{x_i}}{\sum_{i=1}^{n} e^{x_i}}$

1) Compute Softmax o/p
2) Compute the Jacobian

$s_1 = \frac{e^{x_1}}{\sum_{i=1}^{3} (e^{x_1} + e^{x_2} + e^{x_3})}$

$s_1 = \frac{e^2}{(e^2 + e^1 + e^{0.1})}$

$= \frac{7.3}{7.3 + 2.7 + 1.1}$

$= \frac{7.3}{11.1} \Rightarrow 0.66$

$s_2 = \frac{e^{x_2}}{11.1} = \frac{e^1}{11.1} = \frac{2.7}{11.1} \Rightarrow 0.24$

$s_3 = \frac{e^{x_3}}{11.1} = \frac{1.1}{11.1} = 0.099 \Rightarrow 0.1$

$s = \begin{bmatrix} s_1 \\ s_2 \\ s_3 \end{bmatrix} = \begin{bmatrix} 0.66 \\ 0.24 \\ 0.1 \end{bmatrix}$

$\frac{ds_1}{dx_1} = s_1(1 - s_1)$

$J_{ij} = \frac{\partial s_i}{\partial x_j}$

$J = \begin{bmatrix} \frac{ds_1}{dx_1} & \frac{ds_1}{dx_2} \end{bmatrix}$

Case 1: $i = j$ (Diagonal element)

$\frac{ds_i}{dx_i} = s_i(1 - s_i) =$

if j Case 2: $\frac{ds_i}{dx_j} = -s_i s_j$

$\frac{ds_1}{dx_1} = s_1(1 - s_1)$

$= 0.66(1 - 0.66)$

$= 0.22.$

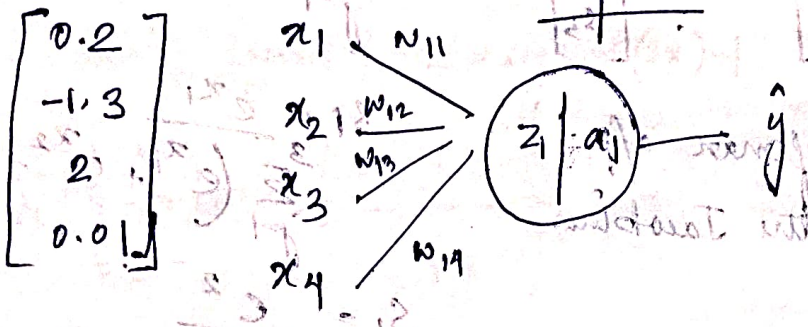
$\frac{ds_1}{dx_j} = -s_1 s_2$

$= -0.66 \times 0.24 = -0.15$

$J = \begin{bmatrix} 0.22 & -0.15 & -0.066 \\ -0.15 & 0.18 & -0.02 \\ -0.06 & 0.02 & 0.09 \end{bmatrix}$

2.C. Saturates in the case of Sigmoid

Lab 2



Layer 1

$w_{11}x_1 + w_{12}x_2 + w_{13}x_3 + w_{14}x_4$

$w_{11} = 0.001$
$w_{12} = 0.01$
$w_{13} = -0.005$
$w_{14} = -1.2$

$z_1 = (0.001 \times 0.2) + (0.01 \times -1.3) + (-0.005 \times 2) + (-1.2 \times 0.01)$

$= 0.0002 + (-0.013) + (-0.01) + (-0.012)$

$= -0.024 - 0.0348$

$a_1 = \text{softmax}(z)$

$z = \begin{bmatrix} -0.024 \\ -0.0348 \end{bmatrix}$

$\begin{bmatrix} \frac{e^{-0.024}}{e^{-0.024} + e^{-0.0348}} \\ z_1 \end{bmatrix}$

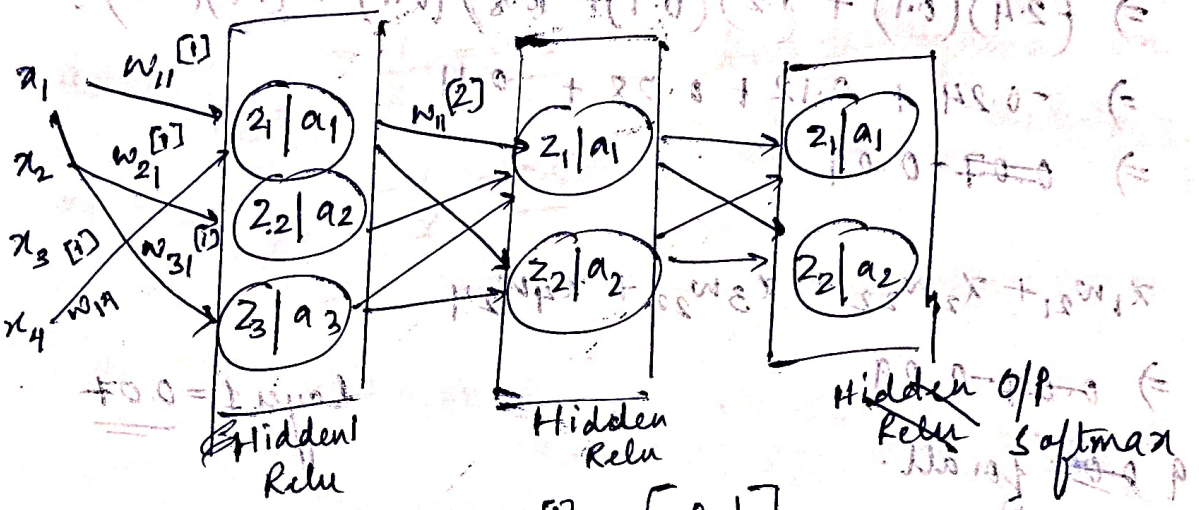
$e^{0.2}$ / $e^{-1.3} + e^{2} + e^{0.01}$

$= \frac{1.22}{0.27 + 0.27 + 7.3 + 1.1}$

$$\begin{bmatrix} 0.2 \\ -1.3 \\ 1.2 \\ 0.01 \end{bmatrix} \begin{bmatrix} 0.001 \\ 0.01 \\ -0.005 \\ -1.2 \end{bmatrix} + \begin{bmatrix} 0.2 \\ -1.3 \\ 2 \\ 0.01 \end{bmatrix} \begin{bmatrix} 0.001 & 0.01 & -0.005 & -1.2 \end{bmatrix}$$

$$= 0.002 + 0 - 0.013 + 0.01 - 0.012$$

$$= -0.0348$$



$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} -2.4 \\ 1.2 \\ -0.8 \\ 1.1 \end{bmatrix}$$

$$w_{11}^{[1]} = \begin{bmatrix} 0.1 \\ 0.1 \\ \vdots \\ 0.1 \end{bmatrix}$$

$w_{21}^{[1]}$

$$x_1 = \frac{-2.4 + 2.4}{2} = 0$$

$$x_2 = \frac{1.2 + 1.2}{2} = \frac{2.4}{2} = 1.2$$

$$x_3 = \frac{-0.8 + 0.8}{2} = 0$$

$$x_4 = \frac{1.1 + 1.1}{2} = \frac{2.2}{2} = 1.1$$

$$z_1 = w_{11} x_1 + x_1 w_{21} + x_1 w_{31} + x_4 w_{14}$$

$$= (0.1)(-2.4) + (2.4)(0.1) + (-2.4)(0.1) + (1.1)(0.1)$$

$x_1 w_{11} + x_1 w_{21} + x_1 w_{31} + x_2 w_{12} + x_2 w_{22} + x_3 w_{32}$

$+ x_3 w_{13} + x_3 w_{23} + x_3 w_{33} + x_4 w_{14} + x_4 w_{24}$

$+ x_4 w_{34}$

$\Rightarrow x_1 w_{11} + x_2 w_{12} + x_3 w_{13} + x_4 w_{14}$

$\Rightarrow (2.4)(0.1) + (1.2)(0.1) + (0.8)(0.1) + (1.1)(0.1)$

$\Rightarrow -0.24 + 0.12 + 0.08 + 0.11$

$\Rightarrow -0.09$

$\Rightarrow x_1 w_{21} + x_2 w_{22} + x_3 w_{23} + x_4 w_{24}$

$\Rightarrow -0.09$

-0.09 for all.

Layer 1 = 0.07

$a_1 = \text{Softmax}$

$= \frac{e^{-0.09}}{e^{-0.09} + e^{-0.09}}$

$= \frac{1.09}{-0.09 + 0.09}$

$a_1 = \text{Relu}(0.09)$

$= 0$

$\frac{(x+|x|)}{2}$

$\frac{-0.09 + 0.09}{2}$

Layer 1

$z_1 = 0 - 0.09 \qquad a_1 = 0$

$z_2 = -0.09 \qquad a_2 = 0$

$z_3 = -0.09 \qquad a_3 = 0$

Layer 2

0 for all z_1 and z_2 values

New x values are ~~OR~~ a_1, a_2 & $a_3 = 0$.

Hidden relu = 0.

Layer 3

$$a_1 = \frac{e^0}{e^0 + e^0} = \frac{1}{2} = 0.5$$

$$a_2 = \frac{e^0}{e^0 + e^0} = \frac{1}{2} = 0.5$$

doing the softmax on a_1 & a_2

Representation power of NN

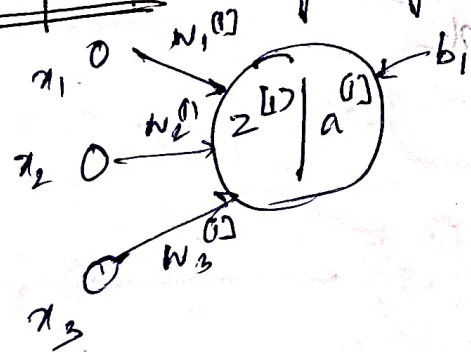
- NN with atleast one hidden layer are universal approximators
- any continuous fn, $f(x)$, and some $\epsilon > 0$, there exists a neural network, $g(x)$, with at least one hidden layer (with some non-linearity), such that

$$\forall x, |f(x) - g(x)| \leq \epsilon$$

both functions mimic hence difference is zero.

Neural Network are non-convex

Perceptron – single layer NN.



$$a^{[1]} = g(z^{[1]})$$

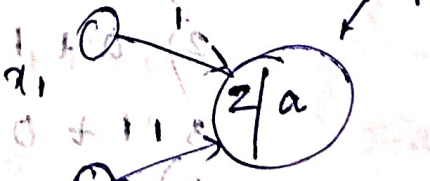
$$z^{[1]} = w_{11}^{[1]} x_1 + w_{12}^{[1]} x_2 + w_{13}^{[1]} x_3 + b_1^{[1]}$$

$$a^{[1]} = g(z^{[1]})$$

$$g = \begin{cases} 1 & \text{if } z > 0 \\ 0 & \text{if } z \leq 0 \end{cases}$$

AND

x_1	x_2	y
0	0	0
0	1	0
1	0	0
1	1	1



$w_{11} = 1$

$w_{12} = 1$

$w_{11} x_1 + w_{12} x + (-1)$

$0 + 0 - 1 = -1$

$g(-1) = \frac{e^{-2}}{e}$

$= 0$

for $x_1=0$ $x_2=1$

$0 + 1 - 1 = 0 \quad g(z)$

for $x_1=1$, $x_2=0$ if bias is 0 $= 1$ so 1

$= 1+0-1 = 0 \quad g(z)$

if bias is 0 $\Rightarrow 1+0+0$

$\geq 1 \quad g(z) = 1$

for $x_1=1$, $x_2=1$

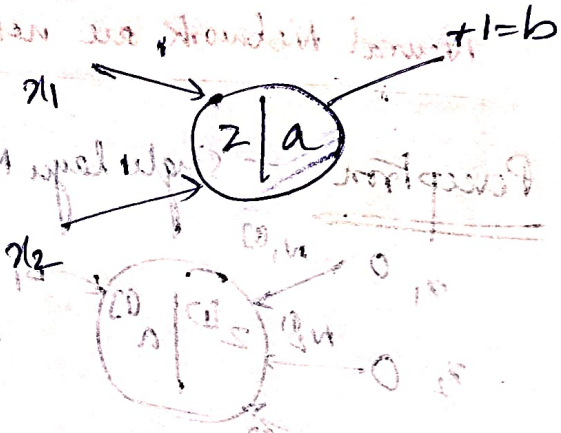
$\Rightarrow 0+1+1-1 \Rightarrow 2-1$

$\Rightarrow 1$

so $g(z) = 1 \quad z > 0$.

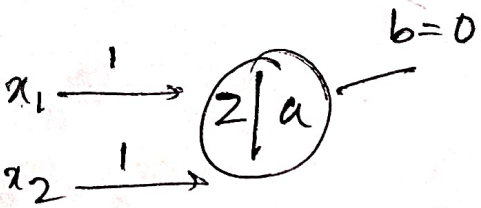
OR

x_1	x_2	y
0	0	0
0	1	1
1	0	1
1	1	1



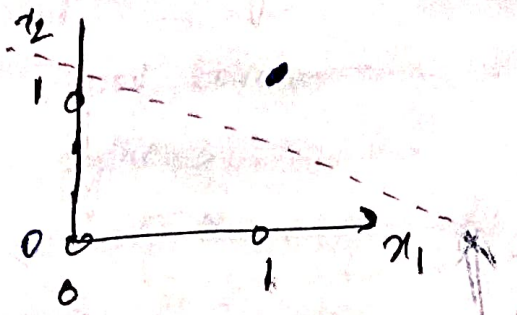
XOR

x_1	x_2	y
0	0	0
0	1	1
1	0	1
1	1	0

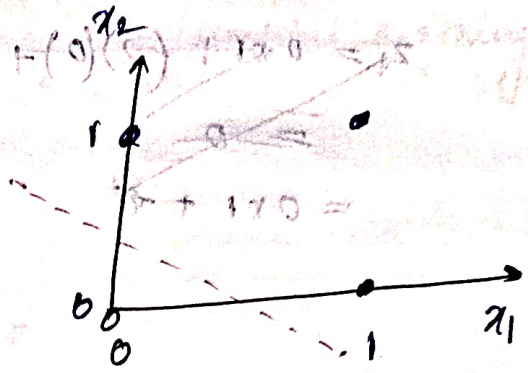


1) $0+0+0=0 \quad g(z)=0$
2) $0+1+0=1 \quad g(z)=1$
3) $1+0+0=1 \quad g(z)=1$
4) $1+1+0=2 \quad g(z)=1$

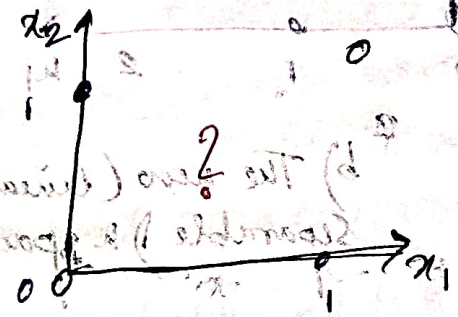
Decision Boundaries



a) x_1 AND x_2

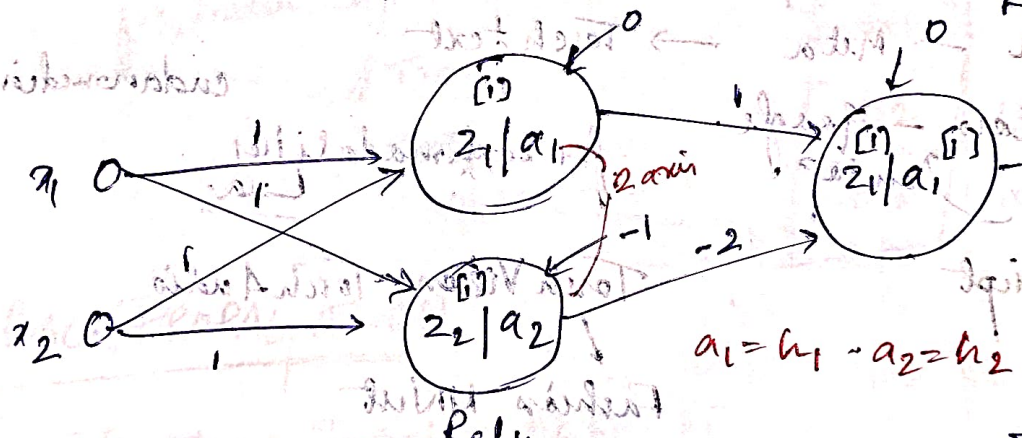


b) x_1 OR x_2



c) x_1 XOR x_2

0 — zero
1 — one



$a_1 = h_1 \cdot a_2 = h_2$

x_1	x_2	y
① 0	0	0
② 0	1	1
③ 1	0	1
④ 1	1	0

$z_1 = w_{11}x_1 + w_{12}x_2 + b_1$ = ~~1×1~~ $0+0+0=0$ ~~=0~~ $g(z)=0$

② $\Rightarrow 0+1+0=1$ $g(z)=1$

③ $\Rightarrow 1+0+0=1$ $g(z)=1$

④ $\Rightarrow 1+1+0=2$ $g(z)=1$ @ a_1

total $z = 4$. $g(z)=1$ $a_1=1$

$z_2 = 0+0-1=-1$ $g(z)=0$

$= 0+1-1=0$ $g(z)=0$

$= 1+0-1=0$ $g(z)=0$

$= 1+1-1=1$ $g(z)=1$

total $z = 1-1=0$ $g(z)=0$ $a_2=0$

Computing the equation through several layers has computational cost ↑

Random initialization is done as 0 will for nets → lead to zero in the equation.

Computational Graph

Net initialization

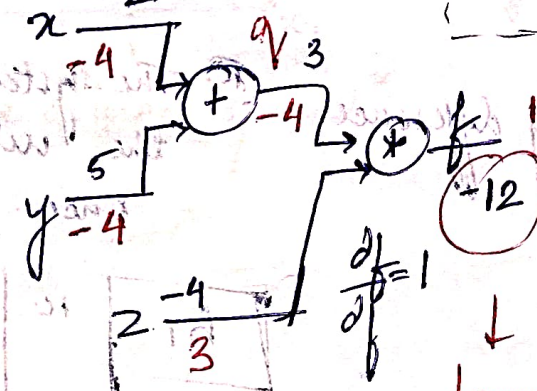
i) $W_{ij} \sim$ uniform $([-0.1, 0.1])$ → uniform distribution

ii) $W_{ij} \sim N\left(0, \sqrt{\frac{2}{n^{[l]} + n^{[l-1]}}}\right)$

→ Xavier init → faster convergence

Let $f = (x+y)z$

multivariable func



z has +ve impact on q. but q has negative impact on f.

forward pass output

Let $q = x + y$

$f = qz \quad \frac{\partial f}{\partial q} = z$

$\frac{\partial q}{\partial x} = 1 \quad \frac{\partial q}{\partial y} = 1 \quad \frac{\partial f}{\partial z} = q$

$\left[\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}, \frac{\partial f}{\partial z}\right]$?

$\Rightarrow \frac{\partial f}{\partial x} = \frac{\partial f}{\partial q} \cdot \frac{\partial q}{\partial x}$

$\Rightarrow \frac{\partial f}{\partial y} = \frac{\partial f}{\partial q} \cdot \frac{\partial q}{\partial y}$

$= z \cdot 1$

$= z \Rightarrow -4$

$= 2 \cdot 1$

$= 2 \Rightarrow -4$

global gradient.

local gradient

output q respect to x — local gradient

then global gradient → value that came to me

- start from back
- put the gradients
- compute the local & global gradients
- multiply them.

29/07/25

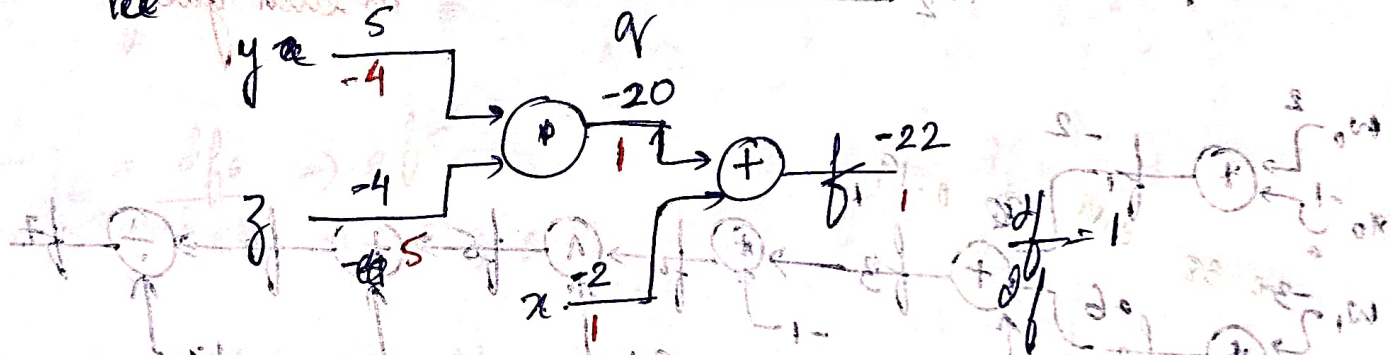
$f(x, y, z) = x + yz$

$x = -2, \quad y = 5, \quad z = -4$

1) Identify additional functions
   1.1 Compute local derivatives.
2) Draw computational graph
3) Perform forward pass
4) Perform backward pass starting from the end of the circuit.

$f = x + yz$

let



let $q = yz \quad f = q + x \longrightarrow$

$\frac{df}{dq} = 1 \quad \frac{df}{dx} = 1$

$\frac{dq}{dy} = z \quad \frac{dq}{dz} = y$

$\Rightarrow -4 \quad \Rightarrow 5$

$\frac{dq}{dx} = 0$

$\frac{df}{dx}, \frac{df}{dy}, \frac{df}{dz}$

$\Rightarrow \frac{df}{dx} = \frac{df}{dq} \cdot \frac{dq}{dx} \Rightarrow 1 \cdot 0$

$\frac{df}{dy} = \frac{df}{dq} \cdot \frac{dq}{dy} \Rightarrow (1)(-4) = -4$

$\frac{df}{dz} = \frac{df}{dq} \cdot \frac{dq}{dz} \Rightarrow (1)(5) = 5$

(*) $\frac{df}{dz}$ - local grad * global grad

$\frac{df}{dy} = \quad , \quad , \quad ,$

$$f(w,x) = \frac{1}{1+e^{-(w_0x_0+w_1x_1+w_2)}}$$

$w_0 = 2$
$x_0 = -1$
$w_1 = 3$
$x_1 = -2$
$w_2 = -3$

~~let $q = w_0x_0 + w_1x_1 + w_2$~~

~~$f = \frac{1}{1+e^{-q}}$~~

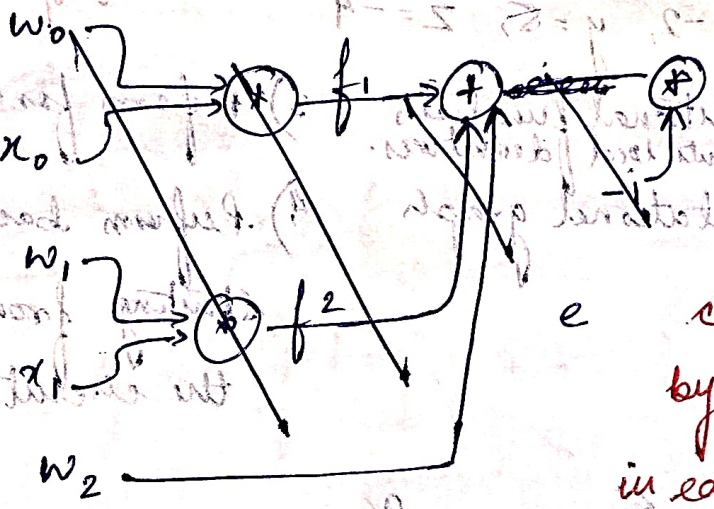
let $f_1 = w_0x_0$ $f_2 = w_1x_1$ $f_3 = f_1 + f_2 + w_2$

$f_4 = -f_3$

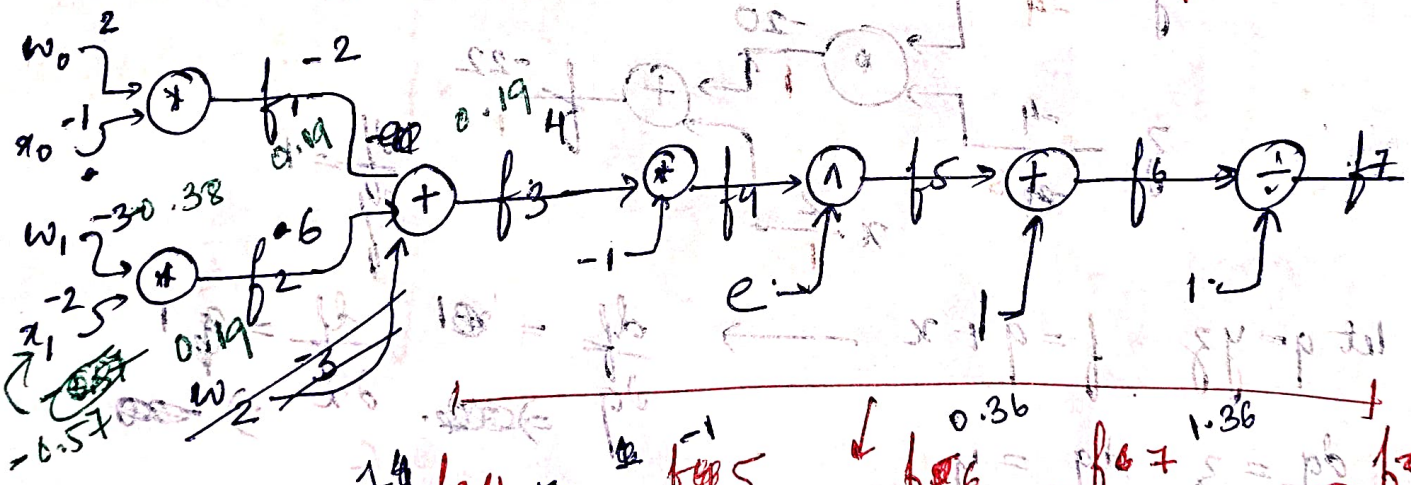
$f_5 = e^{f_4}$

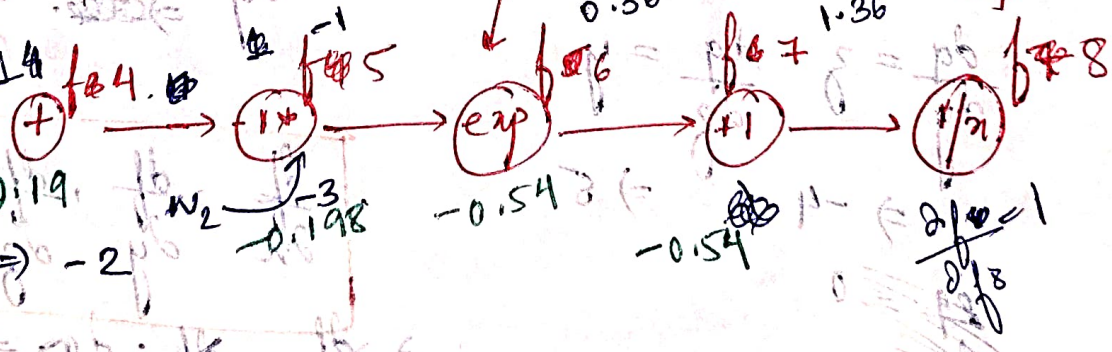
$f_6 = 1 + f_5$

$f_7 = \frac{1}{f_6}$



e can be done by 2 inputs in each gate





$f_1 = 2 \times -1 \Rightarrow -2$

$f_2 = 6$

$f_3 = -2 + (6) + -3 \Rightarrow$ $-2 - 6 - 3$ $-2 + 6 - 3$

$\Rightarrow -2 - 9$

$\Rightarrow -11$

$f_3 = 4$ $\quad f_4 = w_2 + f_3 = -3 + 4 \Rightarrow 1$ $\quad f_5 = -1$

$f_6 = e^{-1} = 0.36$ $\quad f_7 = 1 + 0.36 = 1.36$ $\quad f_8 = \frac{1}{1.36(f_7)} = 0.735$

$f_1 = w_0 x_0$ $\quad f_2 = w_1 x_1$ $\quad f_3 = f_1 + f_2$ $\quad f_4 = f_3 + w_2$

$f_5 = -f_4$ $\quad f_6 = e^{f_5}$ $\quad f_7 = 1 + f_6$ $\quad f_8 = \frac{1}{f_7}$

1 $f_1 \Rightarrow \frac{\partial f_1}{\partial x_0} \Rightarrow w_0 \Rightarrow 2$ $\quad \frac{\partial f_1}{\partial w_0} = x_0 \Rightarrow -1$

2 $f_2 \Rightarrow \frac{\partial f_2}{\partial w_1} = x_1 \rightarrow -2$

$\frac{\partial f_2}{\partial x_1} = w_1 \rightarrow -3$

3 $f_3 = \frac{\partial f_3}{\partial f_1} = 1$ $\quad \frac{\partial f_3}{\partial f_2} = 1$

4 $f_4 = \frac{\partial f_4}{\partial f_3} = 1$ $\quad f_4 \Rightarrow \frac{\partial f_4}{\partial w_2} = 1$

5 $f_5 \Rightarrow \frac{\partial f_5}{\partial f_4} = -1$

6 $f_6 = \frac{\partial f_6}{\partial e^{f_5}} \Rightarrow e^{f_5} \Rightarrow e^{-1} \Rightarrow 0.36$

7 $f_7 = \frac{\partial f_7}{\partial f_6} = 1$

8 $f_8 = \frac{\partial f_8}{\partial (\frac{1}{f_7})} \Rightarrow \frac{\partial f_8}{\partial f_7} \Rightarrow \frac{1}{f_7} \Rightarrow \frac{\partial f_8}{\partial f_7} = \frac{-1}{f_7^2}$

local ones

$\frac{df}{df_8} = \frac{df}{df_8} \times \frac{df_8}{df_7}$ local × global

Global

$\frac{df_8}{df_7} = \text{local} \times \text{global}$

$= \frac{-1}{f_7^2} \times 1 \Rightarrow \frac{-1}{(1.36)^2} \times 1 \Rightarrow -0.54$

$$\frac{\partial f_8}{\partial f_6} = \frac{\partial f_7}{\partial f_6} \times \frac{\partial f_8}{\partial f_7}$$

$$= 1 \times -0.54$$

$$= -0.54$$

$$\frac{\partial f_8}{\partial f_5} = \frac{\partial f_6}{\partial f_5} \times \frac{\partial f_8}{\partial f_6}$$

$$= -0.54 \times 0.36$$

$$= -0.198$$

$$\frac{\partial f_8}{\partial f_4} = \text{local} \times \text{global}$$

$$\Rightarrow \frac{\partial f_5}{\partial f_4} \times \frac{\partial f_8}{\partial f_5}$$

$$\Rightarrow -1 \times -0.198$$

$$\Rightarrow 0.19$$

$$\frac{\partial f_8}{\partial f_3} = \frac{\partial f_4}{\partial f_3} \times \frac{\partial f_8}{\partial f_4}$$

$$= 1 \times 0.19$$

$$= 0.19$$

$$\frac{\partial f_8}{\partial f_2} = \frac{\partial f_3}{\partial f_2} \times \frac{\partial f_8}{\partial f_3}$$

$$= 1 \times 0.19$$

$$\Rightarrow 0.19$$

$$\frac{\partial f_8}{\partial f_1} = \frac{\partial f_2}{\partial f_1} \times \frac{\partial f_8}{\partial f_2}$$

$$=$$

$$\frac{\partial f_8}{\partial f_2} = \frac{\partial f_3}{\partial f_2} \times \frac{\partial f_8}{\partial f_3}$$

$$\Rightarrow 0.19$$

$$\frac{\partial f_4}{\partial f_4} = \frac{\partial f_5}{\partial f_4} \times \frac{\partial f_8}{\partial f_5}$$

$$= 0.19.$$

$$\frac{\partial f_8}{\partial f_3} = \frac{\partial f_4}{\partial f_3} \times \frac{\partial f_8}{\partial f_4}$$

$$= 0.19$$

$$\frac{\partial f_8}{\partial w_1} = \frac{\partial f_2}{\partial w_1} \frac{\partial f_8}{\partial f_2}$$

$$= -2 \times 0.19$$

$$= -.38$$

$$\frac{\partial f_8}{\partial w_2} = \frac{\partial f_2}{\partial w_2} \times \frac{\partial f_8}{\partial f_2} \quad \frac{\partial f_4}{\partial w_2} \times \frac{\partial f_8}{\partial f_4}$$

$$= \times$$

$$\Rightarrow 1 \times 0.19$$

$$= 0.19$$

$$\frac{df_8}{dx_1} = \frac{df_2}{dx_1} \times \frac{df_8}{df_2}$$

$$= w_1 \times 0.19$$

$$= -3 \times 0.19$$

$$= -0.57$$

$$\frac{df_8}{dw_0} = \frac{df_3}{df_1} \times \frac{df_8}{df_3}$$

$$= 1 \times 0.19$$

$$= 0.19$$

$$\frac{df_8}{dw_0} = \frac{df_1}{dw_0} \frac{df_8}{df_{11}}$$

$$= -1 \times 0.19$$

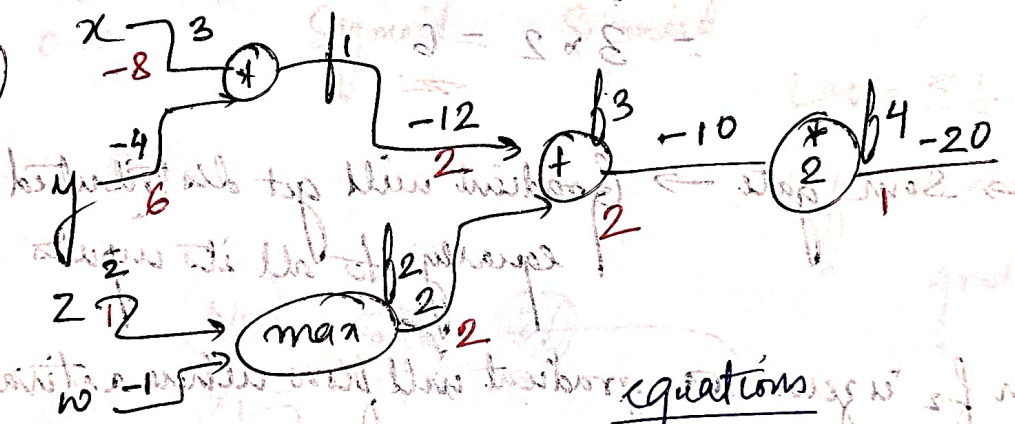
$$= -0.19$$

$$\frac{df_8}{dx_0} = \frac{df_1}{dx_0} \frac{df_8}{df_1}$$

$$= 2 \times 0.19$$

$$= 0.38$$

(2)



equations

$f_1 = xy$

$f_2 = \max(z, w)$

$f_3 = f_1 + f_2$

$f_4 = 2 f_3$

$\frac{df_1}{dx} = y \quad \frac{df_1}{dy} = x$

$= -4 \quad \Rightarrow 3$

$\frac{df_2}{dz} = 1 \rightarrow$ if $z > w$

0 if $z < w$

undefined $z = w$

$\frac{df_2}{dw} = 1$ if $w > z$

0 if $w < z$ undefined $w = z$

$\frac{df_3}{df_1} = 1 \quad \frac{df_3}{df_2} = 1$

$\frac{df_4}{df_3} = 2$

$$\frac{df_4}{df_3} = \frac{df_4}{df_3} \times 1$$

$$= 2$$

$$\frac{df_4}{df_2} = \frac{df_3}{df_2} \frac{df_4}{df_3}$$

$$= 1 \times 2$$

$$= 2$$

$$\frac{df_4}{df_1} = \frac{df_3}{df_1} \cdot \frac{df_4}{df_3}$$

$$= 1 \times 2$$

$$\frac{df_4}{dx} = \frac{df_1}{dx} \cdot \frac{df_4}{df_1}$$

$$= -4 \cdot 2$$

$$= -8$$

$$\frac{df_4}{dy} = \frac{df_1}{dy} \frac{df_4}{df_1}$$

$$= 3 \times 2 = 6$$

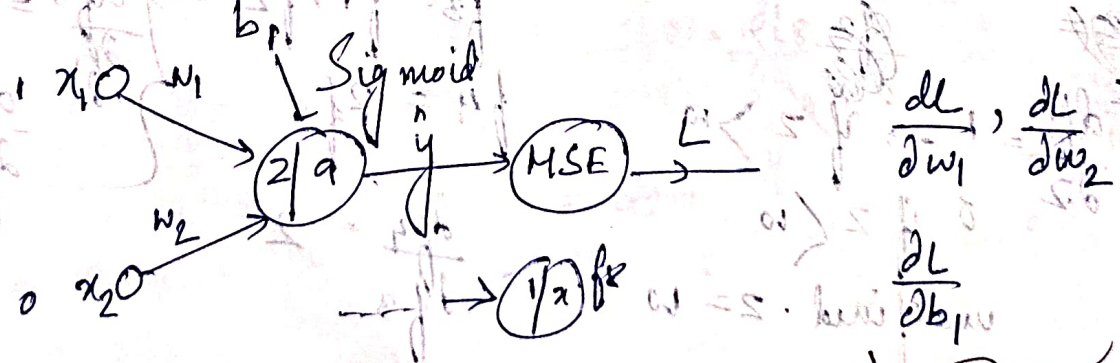
Observations:

~~The gradient~~ → Sum gate → Gradient will get distributed equally to all its inputs

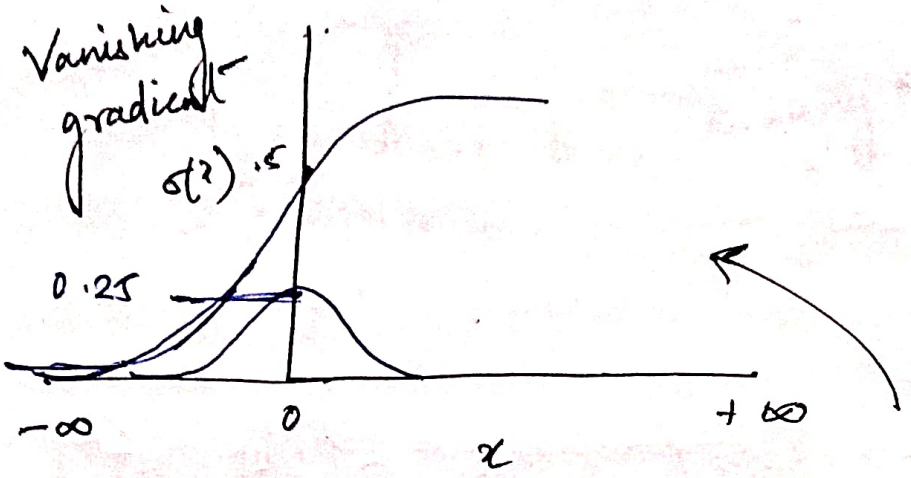
Max gate →

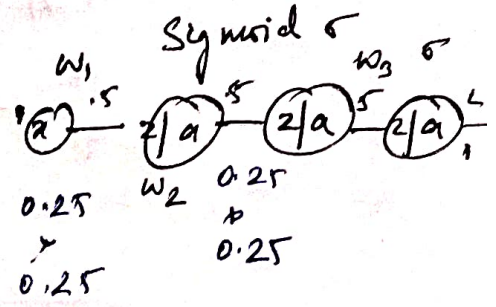
Impact of w on f_2 is zero. gradient will flow where activation values are highs. (Routes the gradient to the highs values)

* gate → values gets swapped. (of the input values)



Vanishing and exploding Gradient Problems





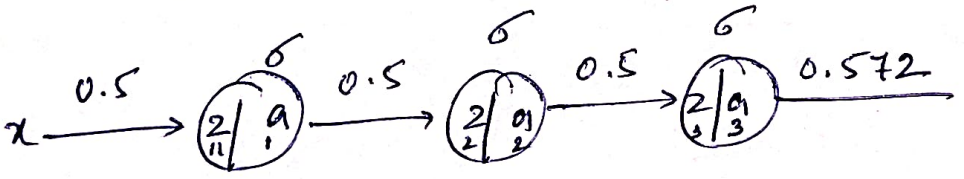
No updates as gradients will be non-existent

gradient is vanished as more decimal number will get multiplied - and no learning will be there.

ReLU ⟶ Exploding gradient problem.

can happen potentially to ReLU

Relu can be solved by → Clipping the gradients ⟶ threshold.



$z_1 = .1 \times 0.5 = 0.5$

$a_1 = \sigma(z_1) = 0.622$

$\frac{1}{1+e^{-z}} = \frac{1}{1+e^{-0.5}} = \frac{1}{1+0.606} = \frac{1}{1.606}$

$L = \frac{1}{2}(a_3 - y)^2$, ~~$\frac{dL}{da_3} = a_3 - y$~~

$= 0.572$

$\frac{dL}{da_3} = \frac{1}{2} \times 2 (a_3 - y)^1$

$= a_3 - y$

~~$a_1 = 0.622$~~

~~$z_2 = 0.5 \times 0.622$~~

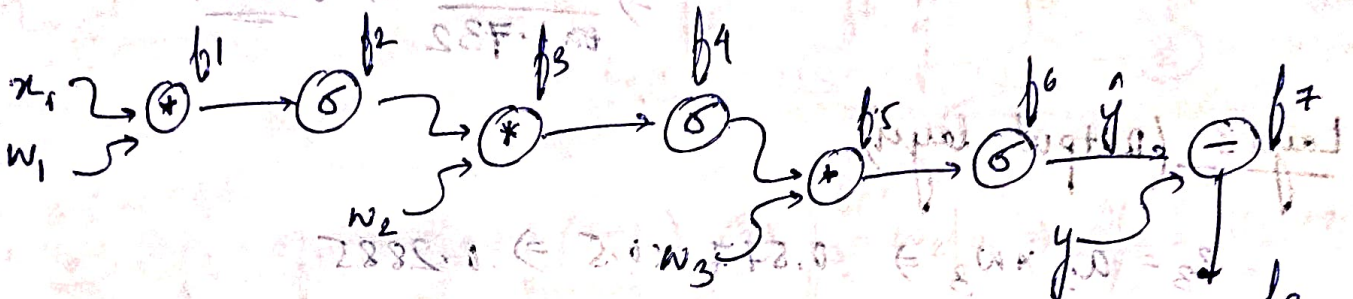
~~$= 0.311$~~

$a_3 - y = 0.572 - 1$

$= -0.428$

~~$\frac{dL}{da_3} = 0.672$~~ 0.

$\frac{dL}{da_3} = -0.428$



$f_1 = x_1 w_1$

$= 1 \times 0.5$

$= 0.5$

$f_2 = \sigma(0.5)$

$= \frac{1}{1+e^{-2}} = \frac{1}{1+0.606}$

$= \frac{1}{1.606}$

$\Rightarrow 0.622$

$f_3 = 0.622 \times 0.5 (w_2)$

$= 0.311$

$f_4 = \sigma(0.311)$

$= \frac{1}{1+e^{-0.311}} = \frac{1}{1+0.732} \Rightarrow \frac{1}{1.732} \Rightarrow 0.577$

$f_5 = 0.577 \times 0.5$

$= 0.2885$

$f_6 = \sigma(f_5)$

$= \frac{1}{1+e^{-0.2885}}$

$= 0.574$

Step 1 Forward Pass

Layer 1

$z_1 = x_1 w_1$

$\Rightarrow 1 \times 0.5 \Rightarrow 0.5$

$a_1 = \sigma(z_1)$

$= \frac{1}{1+e^{-2}} \Rightarrow \frac{1}{1+0.606} \Rightarrow \frac{1}{1.606}$

$\Rightarrow 0.622$

For Vanishing use Sigmoid

Layer 2

$z_2 = a_1 \times w_2 \Rightarrow 0.622 \times 0.5 \Rightarrow 0.311$

$a_2 = \sigma(z_2) \Rightarrow \frac{1}{1+e^{-z}} \Rightarrow \frac{1}{1+e^{-0.311}}$

$\Rightarrow \frac{1}{1.732} \Rightarrow 0.577$

Layer 3 (Output layer)

$z_3 = a_2 \times w_3 \Rightarrow 0.577 \times 0.5 \Rightarrow 0.2885$

$a_3 = \sigma(z_3) \Rightarrow \frac{1}{1+e^{-0.288}} \Rightarrow \frac{1}{1.749} \Rightarrow 0.572$

Loss: $L = \frac{1}{2}(a_3 - y)^2$

$= \frac{1}{2}(0.572 - 1)^2 \Rightarrow 0.092.$ } need to check

Gradient Calculation → $\frac{dl}{da_3} = a_3 - y \Rightarrow 0.572 - 1$

$\Rightarrow -0.428$

Output layer:

$\frac{dl}{dz_3} = \frac{dl}{da_3} \cdot \sigma(z_3) = \frac{dl}{da_3} \cdot a_3(1-a_3)$

$z_3(1-z_3) \Rightarrow 0.2885(1-0.2885)$

$= 0.2885 \cdot 0.7115 \Rightarrow 0.205$

$\frac{dl}{dz_3} \Rightarrow (-0.428)(0.205) \Rightarrow -0.09$

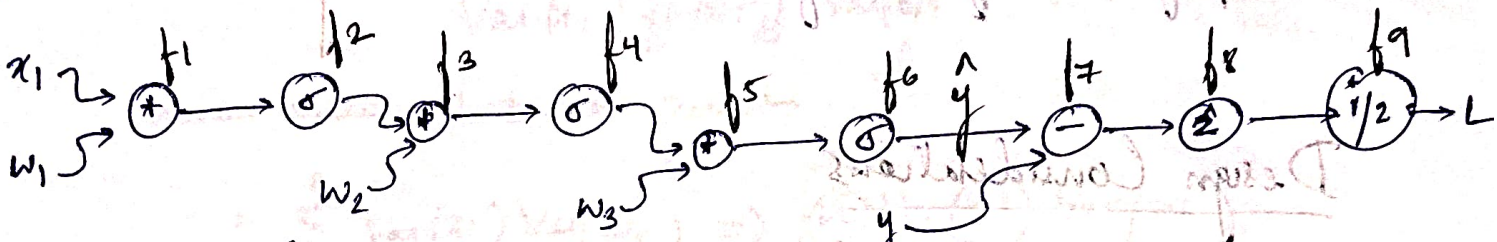
$a_3(1-a_3) = 0.572(1-0.572) \Rightarrow 0.572 \times 0.428$

$\Rightarrow 0.244$

$\frac{dl}{dz_3} = 0.244 \times -0.428 \Rightarrow -0.104.$

2nd layer $\frac{dl}{da_2} = \frac{dl}{dy} \frac{dy}{da_2}$ Use computational graph.

$\frac{df_9}{df_9} = 1$ $\frac{df_9}{df_8} = \frac{dl}{df_8} = 1$ $\frac{df_9}{dl} = \frac{dl}{dl} \times \frac{df_9}{dl}$



① $f_1 = x_1 w_1$ ② $f_2 = \sigma f_1 \Rightarrow \frac{df_2}{df_1} = f_1(1-f_1)$ ③ $f_3 = w_2 f_2 \Rightarrow \frac{df_3}{df_2} = w_2 = 0.5$

$0.5 \times 1 = 0.5$

$\frac{df_1}{dx} = w_1$, $\frac{df_1}{dw_1} = x$

④ $f_4 = \sigma(f_3)$ $\frac{df_4}{df_3} = f_3(1-f_3)$

$\frac{df_3}{dw_2} = f_2 = 0.25$

⑤ $f_5 = w_3 f_4 \Rightarrow \frac{df_5}{df_4} = w_3$, $\frac{df_5}{dw_3} = f_4$

⑥ $f_6 = \sigma(f_5)$ $\frac{df_6}{df_5} = f_5(1-f_5)$

⑦ $f_7 = (f_6 - y)$ $\frac{df_7}{df_6} = 1$ $y = 1$ $f_6 = 0.576$ $= 0.428$ $= 1$

⑧ $f_8 = (f_7)^2 \Rightarrow \frac{df_8}{df_7} = 2(f_7)$

⑨ $f_9 = \frac{f_8}{2} \Rightarrow \frac{df_9}{df_8} = 0.5$

values is the next pag

Backpropagation

$\frac{dl}{dl} = 1$ $\frac{df_9}{df_9} = 1$ $\frac{df_9}{df_8} = \frac{df_9}{df_8} \times \frac{dl}{df_9}$

$= 0.5 \times 1 \Rightarrow 0.5$

$\frac{dl}{df_7} = \frac{dl}{df_8} \times \frac{df_8}{df_7}$

$\Rightarrow 0.25 \times 2(f_7)$ $\Rightarrow 0.5 \times 0.428 \times 2 \Rightarrow 0.428$

$\frac{dl}{df_6} = \frac{dl}{df_7} \times \frac{df_7}{df_6} = 0.428 \times$

Batch norm → Batch Normalization

Normalization per batch.

Normalizing the inputs first

Trained during any I&Op

Introduce before Activation functions

1) nn.Linear( )
2) nn.Batchnorm( )
3) nn.ReLu( ).

Mini Batch GD.

Backpropagation

$\frac{df_9}{df_9} = 1$ $\quad \frac{df_9}{df_8} = 1 \times 0.5 = 0.5$ $\quad \frac{df_9}{df_7} = \frac{df_8}{df_7} \times \frac{df_9}{df_8}$

$= 2(f_7) \times 0.5$

$= 2(0.428) \times 0.5$

$= 0.428$

$f_1 = 0.5$ $\quad f_2 = \sigma f_1 \Rightarrow \sigma f_1 (1 - \sigma f_1)$

$\Rightarrow 0.622(1 - 0.622)$

$\Rightarrow 0.235$

$f_3 = w_2 f_2 \Rightarrow \frac{df_3}{df_2} = 0.5$

$\frac{df_3}{dw_2} = 0.25$

$f_4 \Rightarrow \sigma f_3$

$\frac{df_4}{df_3} = \sigma f_3 (1 - \sigma f_3)$

$= 0.577(1 - 0.577)$

$= 0.244$

$f_5 = w_3 = 0.5$

$f_4 = 0.244$

$f_6 = \sigma f_5 (1 - \sigma f_5) \Rightarrow 0.574(1 - 0.574)$

$= 0.244 \; 0.245$

$f_7 = 0.428$

$\frac{df_9}{df_6} = \frac{df_7}{df_6} \times \frac{df_9}{df_7}$

$1 = 0.428 \times 0.428$

$= 0.1831 \; 0.428$

$f_8 = 2(f_7) = 0.856$

$f_9 = 0.5$

$\frac{df_9}{df_5} = \frac{df_6}{df_5} \times \frac{df_9}{df_6}$

$= 0.245 \times 0.183$

$= 0.0448$

$\simeq 0.045$

$$\frac{df_9}{df_4} = \frac{df_5}{df_4} \times \frac{df_9}{df_5} \Rightarrow 0.5 \times 0.045 \Rightarrow 0.0225$$

$$\frac{df_9}{df_3} = \frac{df_4}{df_3} \times \frac{df_9}{df_4} \Rightarrow 0.244 \times 0.0225 \Rightarrow 0.00549 \simeq 0.0055$$

$$\frac{df_9}{df_2} = \frac{df_3}{df_2} \times \frac{df_9}{df_3} \Rightarrow 0.5 \times 0.0055 \Rightarrow 0.002745$$

$$\frac{df_9}{df_1} = \frac{df_2}{df_1} \times \frac{df_9}{df_2} \Rightarrow 0.235 \times 0.00274 \Rightarrow 0.00064450$$

$$\frac{df_9}{df_6} = \frac{df_7}{df_6} \times \frac{df_9}{df_7} \Rightarrow 1 \times 0.428 \Rightarrow 0.428$$

$$\frac{df_9}{df_5} = \frac{df_6}{df_5} \times \frac{df_9}{df_6} \Rightarrow 0.245 \times 0.428 = 0.10486 \simeq 0.105$$

$$\frac{df_9}{df_4} = \frac{df_5}{df_4} \times \frac{df_9}{df_5} \Rightarrow 0.5 \times 0.105 = 0.0525$$

$$\frac{df_9}{df_3} = \frac{df_4}{df_3} \times \frac{df_9}{df_4} \Rightarrow 0.0525 \times 0.244 \Rightarrow 0.01281 \simeq 0.013$$

$$\frac{df_9}{df_2} = \frac{df_3}{df_2} \times \frac{df_9}{df_3} \Rightarrow 0.5 \times 0.013 \Rightarrow 0.00064$$

$$\frac{df_9}{df_1} = \frac{df_2}{df_1} \times \frac{df_9}{df_2} \Rightarrow 0.235 \times 0.00064 \Rightarrow 0.00150$$

Vanishing Gradient Final value

Exploding Gradient

Taking the weight as 5

$f_1 = x w_1 \Rightarrow 1 \times 5$

$f_2 = ReLu(f_1) \Rightarrow \frac{5+|5|}{2} = 5$

$f_3 = w_2 f_2 \Rightarrow 5 \times 5 = 25$

$f_4 = ReLu(f_3) = \frac{25+25}{2} = 25$

$f_5 = w_3 f_4 \Rightarrow 5 \times 25 \Rightarrow 125$

$f_6 = ReLu(f_5) = 125$

$f_7 = f_6 - y \Rightarrow 125 - 1 \Rightarrow 124.$

$f_8 = (f_7)^2 \Rightarrow (124)^2 \Rightarrow 15376$

$f_9 = 15376 \times 0.5 \Rightarrow 7688$

$\frac{df_9}{df_8} = 0.5$

$\frac{df_2}{df_1} = 1$

$\frac{df_3}{df_2} = w_2 = 5$

$\frac{df_4}{df_3} = 1$

$\frac{df_5}{df_4} = w_3 = 5$

$\frac{df_6}{df_5} = 1 \quad \frac{df_7}{df_6} = 1$

$\frac{df_8}{df_7} = 2f_7 = 2 \times 124 = 248$

Calculations :-

$\frac{df_9}{df_9} = 1 \quad \frac{df_9}{df_8} = \frac{df_9}{df_8} \quad \frac{df_9}{df_8} = 1 \times 0.5 \Rightarrow 0.5$

$\frac{df_9}{df_7} = \frac{df_8}{df_7} \times \frac{df_9}{df_8} \Rightarrow 248 \times 0.5 \Rightarrow 124$

$\frac{df_9}{df_6} = \frac{df_7}{df_6} \times \frac{df_9}{df_7} \Rightarrow 1 \times 124 \Rightarrow 124$

$\frac{df_9}{df_5} = \frac{df_6}{df_5} \times \frac{df_9}{df_6}$

$= 1 \times 124$

$= 124$

$\frac{df_9}{df_4} = \frac{df_5}{df_4} \times \frac{df_9}{df_5} \Rightarrow 5 \times 124$

$\Rightarrow 620$

$$\frac{df_4}{df_3} = \frac{df_4}{df_3} \times \frac{df_3}{df_4} \Rightarrow 1 \times 6.20 \Rightarrow 6.20$$

$$\frac{df_3}{df_2} = \frac{df_3}{df_2} \times \frac{df_2}{df_3} \Rightarrow 5 \times 6.20 \Rightarrow 3100$$

$$\frac{df_2}{df_1} = \frac{df_2}{df_1} \times \frac{df_2}{df_2} \Rightarrow 1 \times 3100 \Rightarrow 3100$$

$$\frac{df_4}{dx} \Rightarrow \cancel{3100 \times} \frac{df_1}{dx} \times \frac{df_4}{df_1} \Rightarrow 5 \times 3100 \Rightarrow 15500$$

Final Exploding Gradient.

Batch Norm Continued

Scale and shift operation on top of variance and other operations

↳ mean of gaussian is something else

Shifting of space might happen wrongly. The datapoint might shift somewhere else.

- Normalize each layer
- Apply to $z^{[l]}$
- scale and shift → Learnable parameters.
  (γ) (β)
  (becomes prominent with each layer)
- $\gamma^{[l]}$ $\beta^{[l]}$.

Example:- $z = [4, 8, 6, 10]$.

- $\mu = 4+8+6+10 \Rightarrow 7$

$$\sigma = \frac{1}{4}\sum_{i=1}^{4}(z_i - \mu)^2$$

$$= 5$$

Convolutional Neural Network

1 sample → 1 image → 64 × 64 × 1 — input layer → image

↓

made of RGB channels → passed as a vector → join and give colour image.

Histogram of Oriented Gradient (HOG)
" " Flows (HOF)
- Optical flow features

} feature extractions.

| 43 | 240 | 1 | 28 | ✗

adjacent pixels will have homogenous property.

feature detector [ ] ← vector.

Convolution Operation.

3	0	1	2	7	4
1	5	8	9	3	1
2	7	2	5	1	3
0	1	3	1	7	8
9	2	1	6	2	8
2	4	5	2	3	9

Input 6×6 image

Superimpose

* Convolution operated

1	0	-1
1	0	-1
1	0	-1

filter or Kernel 3×3

=

-5	-4	0	8
-10	-2	2	3
0	-2	-4	-7
-3	-2	-3	-16

-16

Convolution operation $\Rightarrow (3\times1)+(0\times0)+(1\times-1)+(1\times1)+(5\times0)+(8\times-1)$
$+(2\times1)+(7\times0)+(2\times-1)$

$\Rightarrow 3-1+1+0-8+2+0-2$

$\Rightarrow -5$

$\Rightarrow (0\times1)+(1\times0)+(2\times-1)+(5\times0)+(8\times0)+(9\times-1)+$
$(7\times1)+(2\times0)+(5\times-1)$

$\Rightarrow 0-0-2+5+0-9+7+0-5$

$\Rightarrow -4.$

$\Rightarrow (1\times1)+(2\times0)+(7\times-1)+(8\times1)+(9\times0)+(3\times-1)+(2\times1)$
$(5\times0)+(1\times-1)$

$\Rightarrow 1+0-7+8+0-3+2+0-1$

$\Rightarrow 0$

Similarly $\Rightarrow 2+0-4+9+0-1+5+0-3$

$\Rightarrow 8$

$\Rightarrow 1+0-8+2+0-2+0+0-3 \Rightarrow -10$

$\Rightarrow 5+0-9+7+0-5+1+0-1 \Rightarrow -2$

$\Rightarrow 8+0-3+2+0-1+3+0-7 \Rightarrow 2$

$\Rightarrow 9+0-1+5+0-3+1+0-8 \Rightarrow 3$

$\Rightarrow 2+0-2+0+0-3+4+0-1 \Rightarrow 0$

$\Rightarrow 7+0-5+1+0-1+2+0-6 \Rightarrow -2$

$\Rightarrow 2+0-1+3+0-7+1+0-2 \Rightarrow -4$

$\Rightarrow 5+0-3+1+0-8+6+0-8 \Rightarrow -7$

$\Rightarrow 0+0-3+4+0-1+2+0-5 \Rightarrow -3$

$\Rightarrow 1+0-1+2+0-6+4+0-2 \Rightarrow -2$

$\Rightarrow 3+0-7+1+0-2+5+0-3 \Rightarrow -3$

$\Rightarrow 1+0-8+6+0-8+2+0-9 \Rightarrow -16$

The filter goes horizontally and then slides down by one box and then again moves horizontally

Vertical filter →

1	0	-1
1	0	-1
1	0	-1

→ Vertical edges

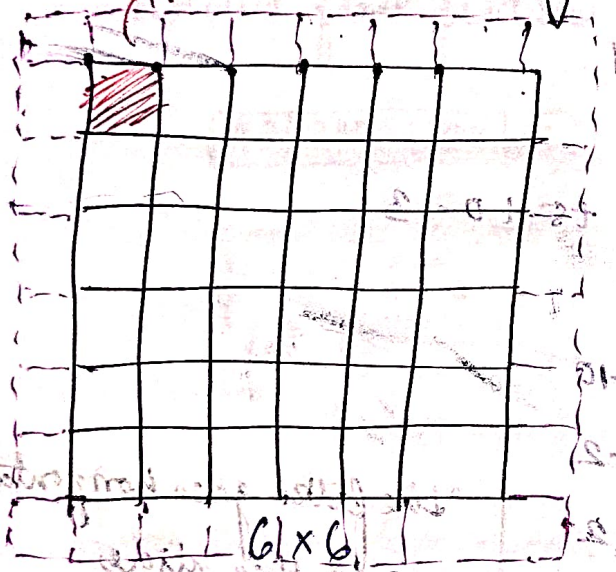
Horizontal filter →

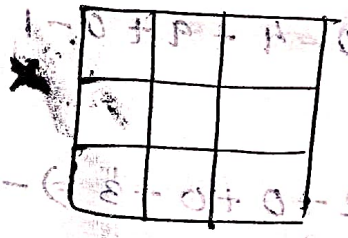
1	1	1
0	0	0
-1	-1	-1

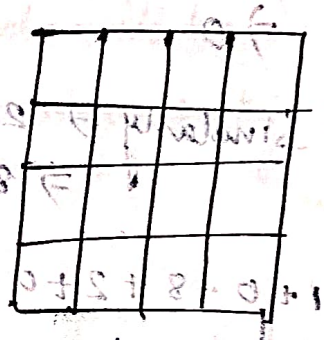
→ Horizontal edges

- edges are taken as pixels and passed it on through SVM.

Padding







n = i/p ; padding = 1 ; p = 1 ; f

$$[n-f+1 \times n-f+1]$$

After applying too many layers it might shrink too much.
and might vanish

Less no of hits → edges
more no of hits → middle

To overcome — shrinking problem, equal attention to all pixels.

$p=1$ $\quad n-f+1 \times n-f+1$

$= (6-3+1) \times (6-3+1)$

$= 4 \times 4$

when $n = 6$ $\quad p = 1$

$f = 3$ $\quad o/p = 6 \times 6$.

for Padding $\quad n+2p-f+1 \times n+2p-f+1$

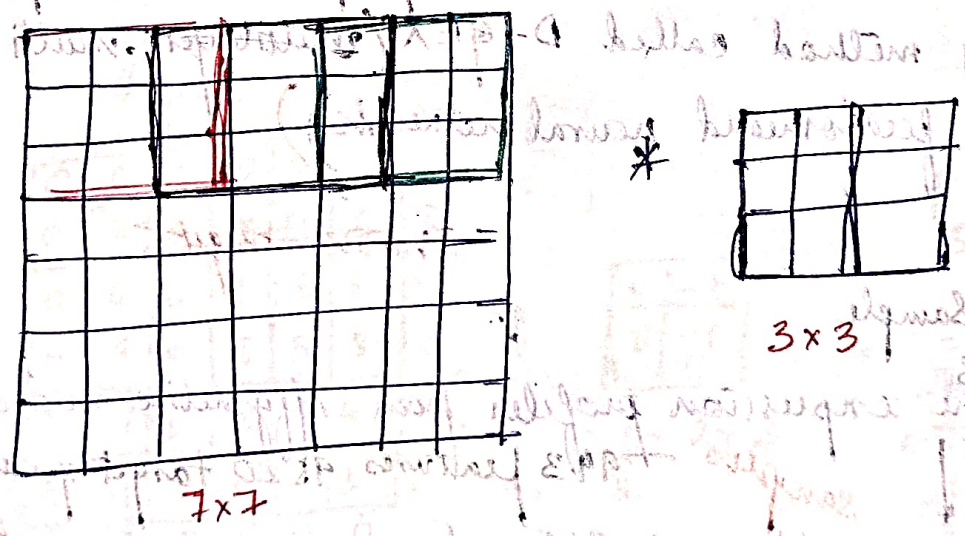
"valid" convolution $\Rightarrow$ i/p size $\neq$ o/p size

"same" convolution $\Rightarrow$ i/p size $=$ o/p size

what should be the value of p to get "same" conv.?

$$p = \frac{f-1}{2}$$

Strided convolutions.



Stride $S = 2$ (hyperparameter)

faster processing $\rightarrow$ leads to shrinked image (even rows x cols)

Padding will solve the issue.

1. GEO
2) GTEx
3. 1000G
4. L1000
   — 110 GB

for training — 943
output — 9520

GEO data → 80% training → 88807
val → 11101
test → 11101

ARIA

last layer is simple linear layer
hence linear activation 2=a or a=2.

19x23 LM_ID, TG_ID } Split

Input: LM_genes — Target_genes: output

pull them out → predict

462 GO samples
1 sample → 22K genes. index into 1000 file.
for every sample

102 genes. input 100 col. 100 total. all the 1000 output.
target

1000 total
100 ← target y

Theory

0	0	0	0	0	0	0
0	2	3	4	4	6	0
0	6	6	9	8	7	0
0	3	4	8	3	8	0
0	7	8	3	6	6	0
0	11	2	1	8	3	0
0	0	0	0	0	0	0

n=5x5

$*$

1	-1
0	-1

f = 2x2

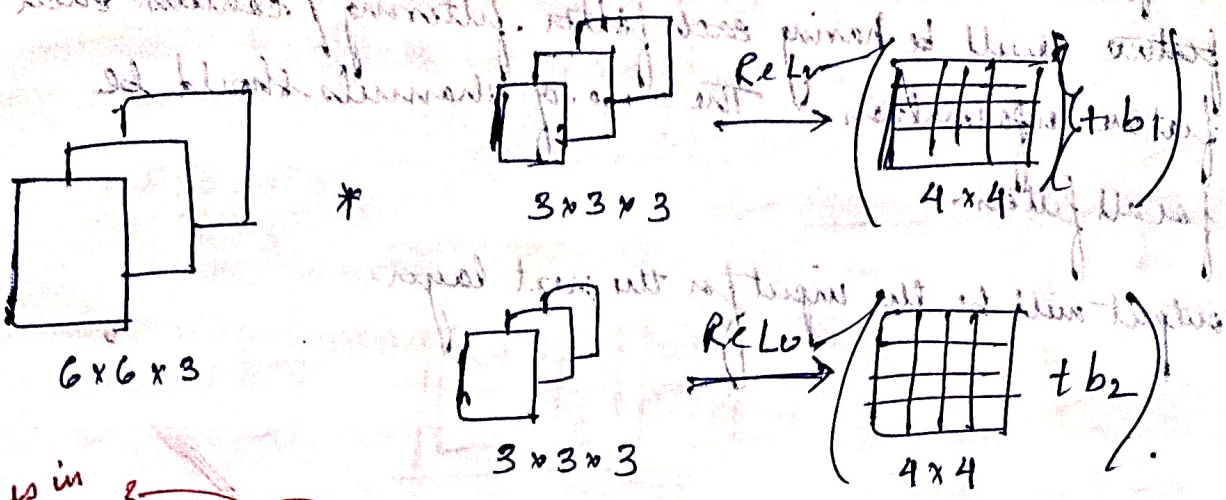
p = 1
S = 2

=

-2	-7	-6
-9	-8	-7
-11	4	-3

= Convolution operation = 0 + 0 + 0 − 2 = −2
= ~~0 + 0 + 0 − 7 = −7~~
= ~~0 + 0 + 0 − 3~~ = 3 0 + 0 + 0 − 7
= 0 + 0 + 0 − 7 = −7
= 0 + 0 + 0 − 6 = −6

−6 − 3 = −9
⇒ 6 − 9 + 0 − 8
⇒ 5



filters in CNN

$z^{[l+1]} = w^{[l+1]} a^{[l]} + b^{[l+1]}$

$a^{[l+1]} = ReLU(z^{[l+1]})$

We apply the filter to a portion of the input and then.

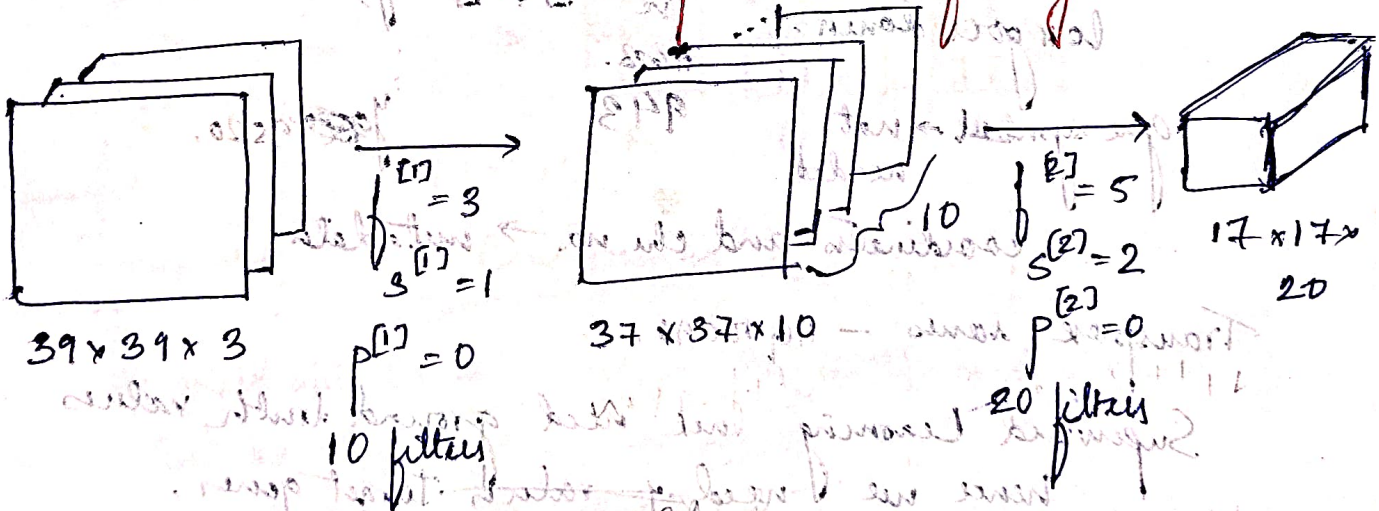
↳ Sparsity of Connections

↳ Weight Sharing

Weights (filter) - Same weight (filter) used across the entire image.

~~Sparsity of Connection~~

↳ These 2 properties make CNN efficient and will make it less prone to overfitting



$$\frac{n+2p-f}{s}+1 = \frac{(39+2(0)-3)}{1}+1$$

$\Rightarrow 36+1$

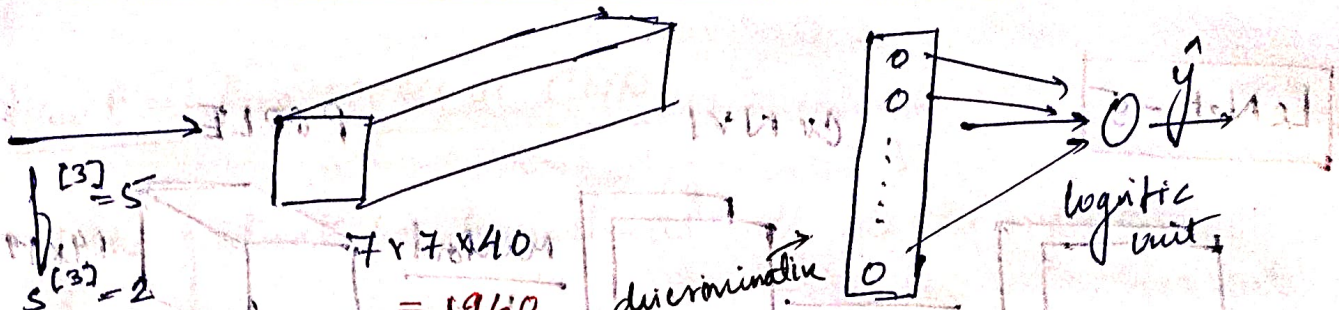
$\Rightarrow 37$

$$\frac{n+2p-f}{s}+1$$

$$= \frac{37-5}{2}+1$$

$$= \frac{32}{2}+1 \Rightarrow 16+1$$

$\Rightarrow 17$



I/P – CONV – CONV – CONV – FC

$$\frac{n+2p-f}{s}+1$$

$$\Rightarrow \frac{17-5+1}{2}$$

$$\Rightarrow \frac{12}{2}+1 \Rightarrow 6+1 \Rightarrow 7$$

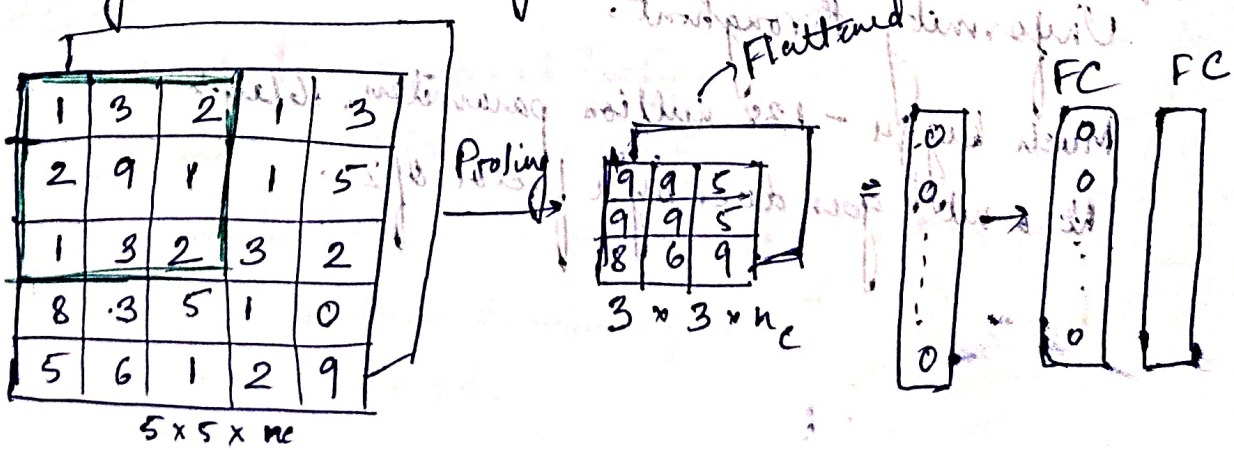
The whole conv is only for one sample.
This is just one row.
(every channel picks up feature)
40 different types of features.
It is combined and flattened and passed through Logits.

10 filters } in one layer of conv net
$3 \times 3 \times 3$ }

How many parameters does this layer have?

$3 \times 3 \times 3 = 27 \times 10$ filters $\Rightarrow$ 270 + 10 biases
$\Rightarrow$ 280

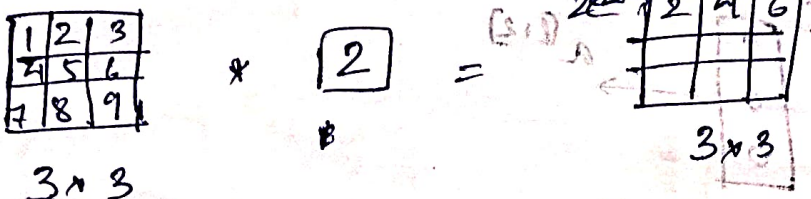
1. Conv. layer
2. Pooling layer — Max Pooling (Most commonly used) computationally less expensive.
   — Avg Pooling
3. Fully Connected (FC) layer.

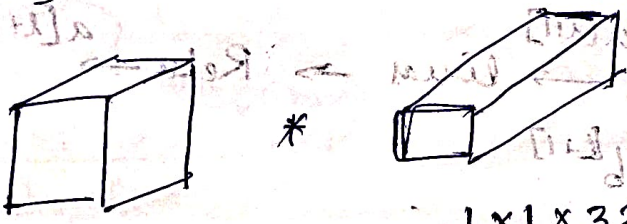


$$a^{[l+2]} = ReLU(z^{[l+2]} + a^{[l]})$$

$64 \times 1 \qquad 32 \times /+W \cdot 64 \times 1$

1 × 1 Convolution :

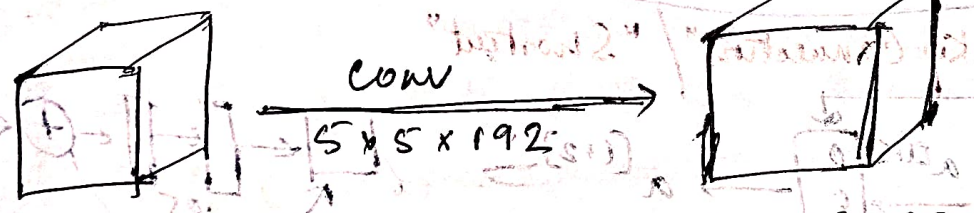




for one output (1 MLP)

Also called Network in a network → if this 1 × 1 × 32 is overlapped on 6 × 6 × 32 then it just becomes a MLP.

contains weights



filters = 32

Same

Total operations = (mult.)

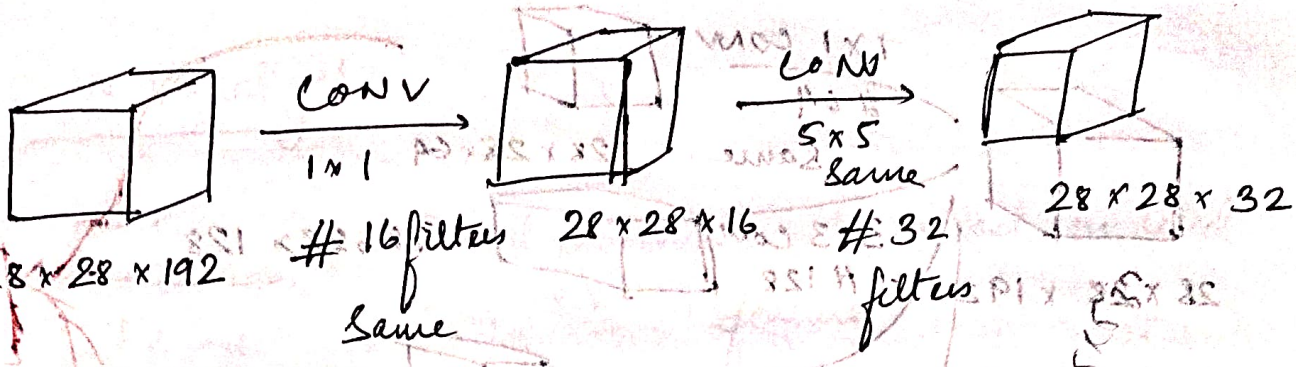
28 × 28 × 192 each filter → 5 × 5 × 192

→ 4800

32 filters → 4800 × 32

→ 153600 × 28 × 28

→ 120,422,400



$28 \times 28 \times 192 = 150528$

$150528 \times 16 \text{ filters} = 2408448$

$2408448 \times 5 \times 5 \times 32 = 1926758400$

$5 \times 5 \times 32 \times 28 \times 28 \xrightarrow{\times 16} 627,200$

$10,035,200$

Bottleneck layer

from 192 to 16 to 32.

$2408448 + 627200$

$2408448 + 10035200 = 12443648$

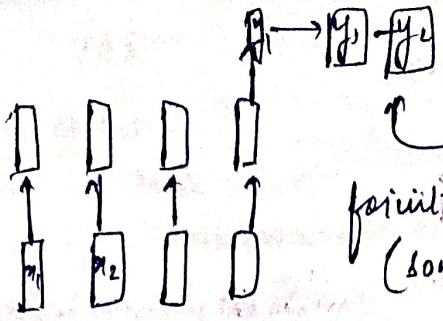
1) Modify height width
   - → Padding
   - → Striding, Pool

2) Modify channels
   - → filters
   - → 1×1 CONV

INCEPTION NETWORK

Just put all filters and let the machine learn from it.

Inception Module → Inception module



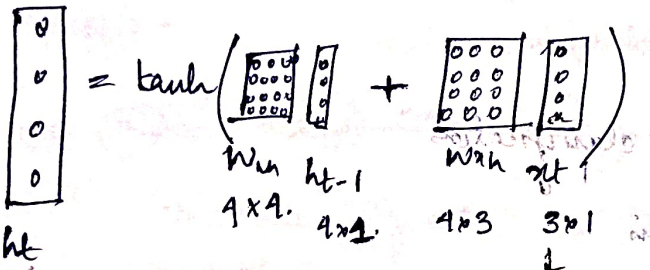
→ sentence

finitial blocks it is optional.

(sometimes we might not even want any output)

h_t is a function $= f(h_{t-1}, x_t)$

$= \tanh(W_{hh} h_{t-1} + W_{xh} x_t)$

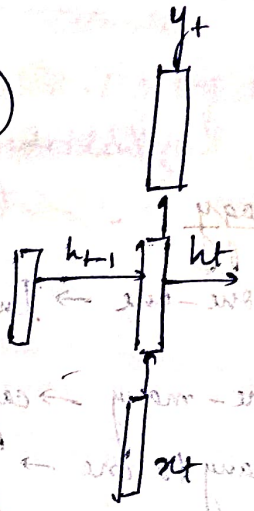


W_{hh} 4×4 · h_{t-1} 4×1 · W_{xh} 4×3 · x_t 3×1

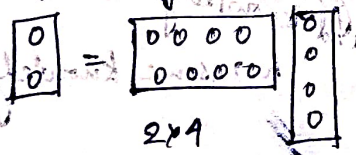
h_t 4×1

Fixing input size

(varying input lengths)



$y_t = W_{hy} h_t$



2×4 · h_t 4×1

the size can be decided (column is fixed)

← weight sharing ~~for all~~

(W_{hh} shared across whole thing)

Vanilla RNN

T=2 (two time steps)

$x^{(1)}: x_1 \in R^2 = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$, $x_2 \in R^2 = \begin{bmatrix} -1 \\ 1 \end{bmatrix}$

Input dimensions = 2 $\Rightarrow x_t \in R^2$

hidden state dim = 3 $\Rightarrow h_t \in R^3$

$$h_0 = \begin{bmatrix}0\\0\\0\end{bmatrix}_{3\times1}, \quad W_{xh} = \begin{bmatrix}0.5 & -0.3\\ 0.8 & 0.2\\ 0.1 & 0.4\end{bmatrix}_{3\times2}, \quad W_{hh} = \begin{bmatrix}0.1 & 0.4 & 0.0\\ -0.2 & 0.3 & 0.2\\ 0.05 & -0.1 & 0.2\end{bmatrix}_{3\times3}$$

$y_t \in R^2$

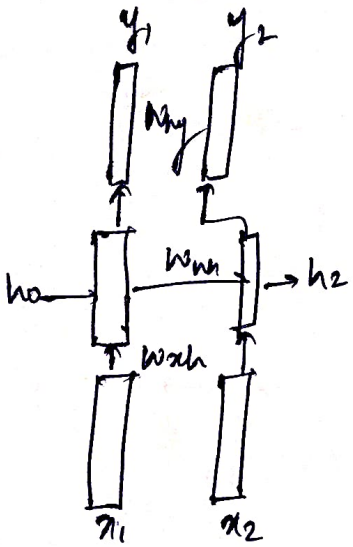
$h_t = 3\times1$

$x_t = 2\times1$

$y_t = W_{hy} h_t \Rightarrow 2 = [2\times3]\,(3\times1) \rightarrow h_t$

$$\begin{bmatrix}1.0 & -1.0 & 0.5\\ 0.5 & 0.5 & -0.5\end{bmatrix}$$

$h_t = \tanh\left(W_{hh} h_{t-1} + W_{xh} x_t\right)$



$h_1, y_1 ?$ $\quad h_t = \tanh(W_{hh} h_{t-1} + W_{xh} x_t)$

$h_2, y_2 ?$ $\quad y_t = W_{hy} h_t$

Time step = 1, $t = 1$

$h_t = \tanh(W_{hh} h_{t-1} + W_{xh} x_t)$

$h_1 = \tanh(W_{hh} h_0 + W_{xh} x_1)$

$$= \tanh\left(\begin{bmatrix}0\\0\\0\end{bmatrix} + \begin{bmatrix}0.5-0.3(2)\\ 0.8+0.2(2)\\ 0.1+0.4(2)\end{bmatrix}\right)$$

$$= \tanh\begin{bmatrix}-0.1\\ 1.2\\ 0.9\end{bmatrix} = \begin{bmatrix}-0.099\\ 0.833\\ 0.716\end{bmatrix}$$

$\frac{e^2 - e^{-2}}{e^2 + e^{-2}}$

$\frac{0.9 - 1.105}{0.9 + 1.105} = \frac{-0.201}{2.009}$

$\frac{3.3 - 0.3}{0.3 + 3.3} = \frac{3}{36}$

Practical → time steps as input.

$$y_1 = W_{hy} \cdot h_1$$

$$= \begin{bmatrix} 1.0 & -1.0 & 0.5 \\ 0.5 & 0.5 & -0.5 \end{bmatrix} \times \begin{bmatrix} 0.099 \\ 0.833 \\ 0.716 \end{bmatrix}$$

$$= \begin{matrix} 1.0 \times 0.099 + -1.0 \times 0.833 + 0.5 \times 0.716 \\ 0.5 \times -0.099 + 0.5 \times 0.833 + -0.5 \times 0.716 \end{matrix}$$

$$= \begin{bmatrix} 0.099 + -0.833 + 0.358 \\ -0.0495 + 0.4165 + -0.358 \end{bmatrix}$$

$$= \begin{bmatrix} -0.574 \\ 0.009 \end{bmatrix}$$

$$h_2 = \tanh\left(W_{hh} h_1 + W_{xh} x_2\right)$$

$$= \tanh\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} + \begin{bmatrix} -0.5 & -0.3 \\ -0.8 & +0.2 \\ -0.1 & +0.4 \end{bmatrix}\right)$$

$$= \tanh\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \begin{matrix} -0.099 \\ 0.833 \\ 0.716 \end{matrix} + \begin{bmatrix} -0.8 \\ -0.6 \\ 0.3 \end{bmatrix}\right)$$

$$\Rightarrow \tanh\begin{bmatrix} -0.8 \\ -0.6 \\ 0.3 \end{bmatrix} = \begin{matrix} -0.66 \\ -0.5 \\ 0.29 \end{matrix} \quad \begin{bmatrix} -0.899 \\ 0.233 \\ 0.416 \end{bmatrix}$$

$$\Rightarrow -0.71$$

$$\tanh\left(\begin{bmatrix} 0.1 & 0.4 & 0.0 \\ -0.2 & 0.3 & 0.2 \\ 0.05 & -0.1 & 0.2 \end{bmatrix}\begin{bmatrix} -0.099 \\ 0.833 \\ 0.716 \end{bmatrix} + \begin{bmatrix} -0.8 \\ -0.6 \\ 0.3 \end{bmatrix}\right)$$

$$\Rightarrow \begin{bmatrix} -0.0099 + 0.332 + 0.000 \\ 0.0198 + 0.2499 + 0.1432 \\ -0.0049 + -0.0833 + 0.1432 \end{bmatrix} + \begin{bmatrix} -0.8 \\ -0.6 \\ 0.3 \end{bmatrix}$$

$$\Rightarrow \begin{bmatrix} 0.3221 & -0.8 \\ 0.4129 & -0.6 \\ 0.055 & +0.3 \end{bmatrix} \Rightarrow \begin{bmatrix} -0.4779 \\ -0.1871 \\ 0.355 \end{bmatrix}$$

$\Rightarrow$

element wise

$$c_t = f_t \odot c_{t-1} + i_t \odot g_t$$

$$h_t = o_t \odot \tanh(c_t)$$ current/new c_t

Memory controlled by gates.

Gates will either allow or not allow.

forget gate — $f_t = \sigma(W_{hf} h_{t-1} + W_{xf} x_t)$

input gate — $i_t = \sigma(W_{hi} h_{t-1} + W_{xi} x_t)$

output gate — $o_t = \sigma(W_{ho} h_{t-1} + W_{xo} x_t)$

Sigmoid (smooth gates) — it is a boolean thing.

not a gate — $g_t = \tanh(W_{hg} h_{t-1} + W_{xg} x_t)$

i_t is a vector with ranges from 0 to 1.

g_t is not a gate, but is an update term.

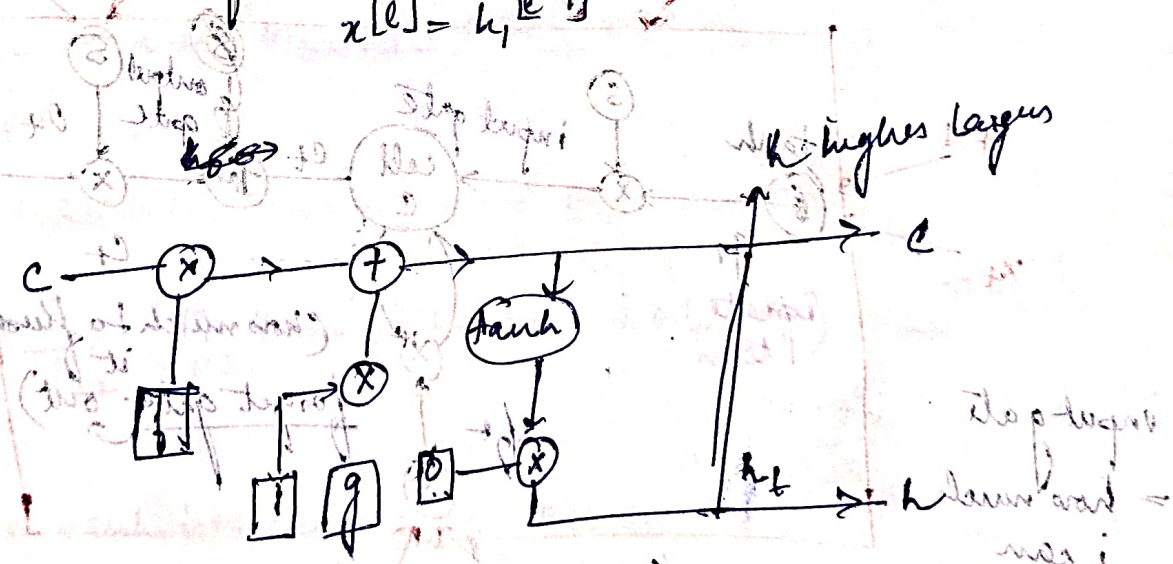
because of gates some or few will be allowed through the gate.

it can keep the entire thing or flush out the entire thing.

c_t = previous memory.

h_t is gated version of c_t

$x[l] = h_1^{[l-1]}$



Stratified modelling (female only / male only) for demographic ft.

Example

$x_t = [0.5, -0.1]$

$h_{t-1} = [0.0, 0.1]$

$c_{t-1} = [0.2, -0.2]$

$W_{xi} = \begin{bmatrix} 0.5 & -0.3 \\ 0.4 & 0.1 \end{bmatrix}$

$W_{hi} = \begin{bmatrix} 0.1 & 0.2 \\ -0.2 & 0.05 \end{bmatrix}$

$W_{xf} = \begin{bmatrix} -0.4 & 0.2 \\ 0.3 & 0.3 \end{bmatrix}$

$W_{hf} = \begin{bmatrix} 0.05 & -0.1 \\ 0.2 & 0.1 \end{bmatrix}$

$W_{xg} = \begin{bmatrix} -0.5 & 0.4 \\ 0.2 & -0.3 \end{bmatrix}$

$W_{hg} = \begin{bmatrix} 0.2 & 0.1 \\ -0.1 & 0.05 \end{bmatrix}$

$W_{xo} = \begin{bmatrix} 0.3 & 0.25 \\ -0.2 & 0.2 \end{bmatrix}$

$W_{ho} = \begin{bmatrix} -0.15 & 0.05 \\ 0.1 & -0.2 \end{bmatrix}$

$h_t = ?$

$c_t = ?$

At t

$$i_t = \sigma(W_{hi} h_{t-1} + W_{xi} x_t)$$

$$= \sigma\left(\begin{bmatrix} 0.15 & 0.05 \\ 0.1 & -0.2 \end{bmatrix}\begin{bmatrix} 0.0 \\ 0.1 \end{bmatrix} + \begin{bmatrix} 0.5 & -0.3 \\ 0.4 & 0.1 \end{bmatrix}\begin{bmatrix} 0.5 \\ -0.1 \end{bmatrix}\right)$$

$W_{hi} h_{t-1}$ calculation wrong. REDO!

$$= \left(0 + 0.005,\ 0 + -0.02\right) + \left(0.25 + -0.03,\ 0.2 + 0.01\right)$$

$$\Rightarrow [0.01, -0.02] + [0.16, -0.14]$$

$$\Rightarrow [0.17, -0.16] \Rightarrow [0.005, -0.02] + [0.22, 0.19]$$

$$\Rightarrow \sigma[0.225, 0.17]$$

$$\frac{1}{1+e^{-2}} \Rightarrow \frac{1}{1+0.78} \Rightarrow \frac{1}{2.24} \Rightarrow 0.44 \quad \frac{1}{1.79} \Rightarrow 0.781 \Rightarrow 0.56.$$

$$\frac{1}{1+0.84} \Rightarrow \frac{1}{1.84} \Rightarrow 0.54$$

$$i_t = [0.78, 0.54] \cdot [0.56, 0.54]$$

$o_t = \sigma(W_{ho} h_{t-1} + W_{xo} x_t)$

= did by mistake in it.

$= [0.005 \quad -0.02] \begin{bmatrix} 0.3 & 0.25 \\ -0.2 & 0.2 \end{bmatrix} \begin{bmatrix} 0.5 \\ -0.1 \end{bmatrix}$

$\Rightarrow [0.15 - 0.025 \quad -0.1 - 0.02]$

$\Rightarrow [0.125 \quad -0.12]$

$\Rightarrow [0.005 \quad -0.002] + [0.125 \quad -0.12]$

$\Rightarrow [0.13 \quad -0.122]$

$\sigma [0.13 \quad -0.122]$

$\Rightarrow \frac{1}{1+e^{-0.13}} = \frac{1}{1+0.87} = \frac{1}{1.87} \Rightarrow 0.53$

$\Rightarrow \frac{1}{1+e^{+0.122}} = \frac{1}{1+1.29} \Rightarrow \frac{1}{2.29} \Rightarrow 0.43$

$o_t \Rightarrow [0.53 \quad 0.43]$

$g_t = \tanh(W_{hg} h_{t-1} + W_{xg} x_t)$

$= \left(\begin{bmatrix} 0.2 & 0.1 \\ -0.1 & 0.05 \end{bmatrix} \begin{bmatrix} 0.0 \\ -0.1 \end{bmatrix} \right) + \left(\begin{bmatrix} -0.5 & 0.4 \\ 0.2 & -0.3 \end{bmatrix} \begin{bmatrix} 0.5 \\ -0.1 \end{bmatrix} \right)$

$= [0 + -0.01 \quad 0 - 0.005] + [-0.25 - 0.04 \quad 0.1 + 0.03]$

$\Rightarrow [-0.01 \quad -0.005] + [-0.29 \quad 0.13]$

$\Rightarrow [-0.35 \quad 0.125]$

$\tanh[-0.35 \quad 0.125]$

$\Rightarrow \frac{e^{-0.35} - e^{+0.35}}{e^{-0.35} + e^{+0.35}} \Rightarrow \frac{0.74 - 1.41}{0.94 + 1.41} \Rightarrow \frac{-0.67}{2.15}$

$\Rightarrow -0.311$

$\Rightarrow \frac{e^{0.125} - e^{-0.125}}{e^{0.125} + e^{-0.125}} = \frac{1.133 - 0.88}{1.133 + 0.88}$

$\Rightarrow \frac{0.253}{2.013} \Rightarrow 0.125$

$\simeq 0.13$

Similarity score → Score.

Alignment score →

Context ki vector.

$h_1 = [1, 0, 1]$ $h_3 = [1, 1, 0]$ $S_{t-1} = [1, 0, 1]$

$h_2 = [0, 1, 1]$

Score fn = dot product.

$c_t = ?$

$$c_t = \sum_{1}^{T} \alpha_{t,j} h_j$$

$$\Rightarrow \alpha_{t,1} h_1 + \alpha_{t,2} h_2 + \alpha_{t,3} h_3$$

$$\alpha_{t,1} = \frac{\exp(\text{score}(S_{t-1}, h_1))}{\exp(\text{score}(S_{t-1}, h_1)) + \exp(\text{score}(S_{t-1}, h_2)) + \exp(\text{score}(S_{t-1}, h_3))}$$

$$= \frac{\exp[1, 0, 1]}{\exp[1, 0, 1] + \exp[1, 1, 1] + \exp[1, 1, 1]}$$

$$= \frac{e^2}{e^2 + e^1 + e^1} \Rightarrow \frac{7.4}{7.4 + 2.7 + 2.7} \Rightarrow \frac{7.4}{12.8} \Rightarrow 0.57$$

$$\alpha_{t,2} = \frac{e^1}{12.8} \Rightarrow \frac{2.7}{12.8} \Rightarrow 0.21 \qquad \alpha_{t,3} = \frac{e^1}{12.8} \Rightarrow 0.21$$

$$c_t = 0.57 + 0.21 + 0.21$$

$$\Rightarrow 099$$

Continuation of prev class

Thinking

① q_1 = Thinking
K_1 = Thinking
v_1 = Thinking

② q_1 = Thinking
K_2 = Machines

Machines

① q_1 = Machines
K_1 = Thinking

② q_2 = Machines
K_2 = Machines

i/p = Playing Outside ⇒ z_1 and z_2

$q_1 = [0.212 \quad 0.04 \quad 0.63 \quad 0.36]^T$

$K_1 = [0.31 \quad 0.84 \quad 0.963 \quad 0.57]^T$

$v_1 = [0.36 \quad 0.83 \quad 0.1 \quad 0.38]^T$

outside

$q_2 = [0.1 \quad 0.14 \quad 0.86 \quad 0.77]^T$

$K_2 = [0.45 \quad 0.99 \quad 0.73 \quad 0.58]^T$

$v_2 = [0.31 \quad 0.36 \quad 0.19 \quad 0.72]^T$

$\sqrt{4} \rightarrow 4$ (dim of k)

$q_1 \times K_1 = 0.06572 + 0.0336 + 0.60664 + 0.2052$

$\Rightarrow 0.91116 / \sqrt{4} \Rightarrow 0.45558 \simeq 0.45 \Rightarrow$ Softmax

$q_1 \times k_2 = 0.0954 + 0.0376 + 0.4599 + 0.2088$

$\Rightarrow 0.8017 \Rightarrow 0.40085 \simeq 0.4 \Rightarrow$

$\frac{e^{0.45}}{e^{0.45} + e^{0.4}}$

$\Rightarrow \frac{1.5}{1.5 + 1.4}$

$\Rightarrow \frac{1.5}{2.9} \Rightarrow 0.5$

$\frac{1.4}{2.9} \Rightarrow 0.48$

Softmax x value $0.5[0.36 \quad 0.83 \quad 0.1 \quad 0.38]$

$\Rightarrow [0.18 \quad 0.415 \quad 0.05 \quad 0.19]$

$0.48[0.31 \quad 0.36 \quad 0.19 \quad 0.72]$

$\Rightarrow [0.148 \quad 0.1728 \quad 0.091 \quad 0.34]$

add

$\Rightarrow [0.328 \quad 0.5878 \quad 0.141 \quad 0.53]$

for 22

$q_2 \times k_2 = [0.031 + 0.1176 + 0.828 + 0.4389]$

$= 1.41155 / \sqrt{4}$

$\Rightarrow 0.70775$

$\simeq 0.7$ softmax $= \frac{2.01}{2.01 + 1.8} \Rightarrow \frac{2.01}{3.81} \Rightarrow 0.52$

$q_2 \times k_2 \Rightarrow [0.45 + 0.1316 + 0.6278 + 0.4466]$

$\Rightarrow 1.251 \Rightarrow 0.62 \Rightarrow$ softmax $\Rightarrow \frac{1.8}{3.81}$

$\Rightarrow 0.472$

Softmax

Softmax x value $\Rightarrow 0.52[0.36 \quad 0.83 \quad 0.1 \quad 0.38]$

$0.4[0.31 \quad 0.36 \quad 0.19 \quad 0.72]$

$\Rightarrow [0.1872 \quad 0.4316 \quad 0.052 \quad 0.1976]$

$+$

$[0.124 \quad 0.144 \quad 0.076 \quad 0.288]$

$\Rightarrow [0.3112 \quad 0.5756 \quad 0.128 \quad 0.4856]$

for this do this with 100

$P_{E_{pos,2i}} = \sin\left(\frac{pos}{100 \cdot (2i/R^q)}\right)$ $P_{pos\,2i+1} = \cos\left(\frac{pos}{100 \cdot (2i/R^q)}\right)$

$i = 0$

$P_{0,0} = \sin\left(\frac{0}{100^{(0/4)}}\right)$ $\sin\frac{0}{100^{0}} \Rightarrow \sin\frac{0}{1}$

$\Rightarrow \sin 0/1$

$\Rightarrow 0.0$

$P_{0,1} = \cancel{\sin 0}\ \cos\left(\frac{0}{100^{1/4}}\right)$

$\Rightarrow \cos\frac{0}{1.2} \Rightarrow 1$

$i = 1$

$P_{0,2} = \sin\left(\frac{0}{100^{2/4}}\right)$

$\Rightarrow \sin\frac{0}{10}$

$\Rightarrow 0$

$P_{0,3} = \cos\frac{0}{100^{3/4}}$

$\Rightarrow \cos\frac{0}{31} = 1$

Position 1

$i = 0$

$P_{1,0} = \sin\frac{1}{100^{0/4}} \Rightarrow \sin\frac{1}{1} = 0.017$

$P_{1,1} = \cos \frac{1}{100^{0/4}} \Rightarrow \cos \frac{1}{1.2} \Rightarrow 0.999$

$i = 1$

$P_{1,2} = \sin \frac{1}{100^{2/4}} \Rightarrow \sin \frac{1}{10} \Rightarrow 0.0017$

$P_{1,3} = \cos \frac{1}{100^{3/4}} \Rightarrow 0.999.$

Position 2

$i = 0$

$P_{2,0} = \frac{\sin 2}{100^{0/4}} = \sin \frac{2}{1} = 0.034$

$P_{2,1} = \cos \frac{2}{100^{0/4}} = \cos \frac{2}{1.2} \Rightarrow 0.999$

$P_{2,2} = \sin \frac{2}{100^{2/4}} \Rightarrow \sin \frac{2}{10} = 0.00008$

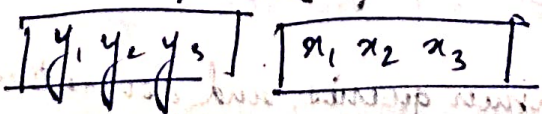
$P_{2,3} = \cos \frac{2}{100^{3/4}} \Rightarrow \sin \frac{2}{1.73} = 0.0006.$

Dot product of initial positional vectors will be high.

P_{500} } dot
P_{501} } prod $= \square$

Continuation

Multihead Attention



in multihead attention the self attention will look at everything but to predict y_4 suppose we need to mask the further ones.