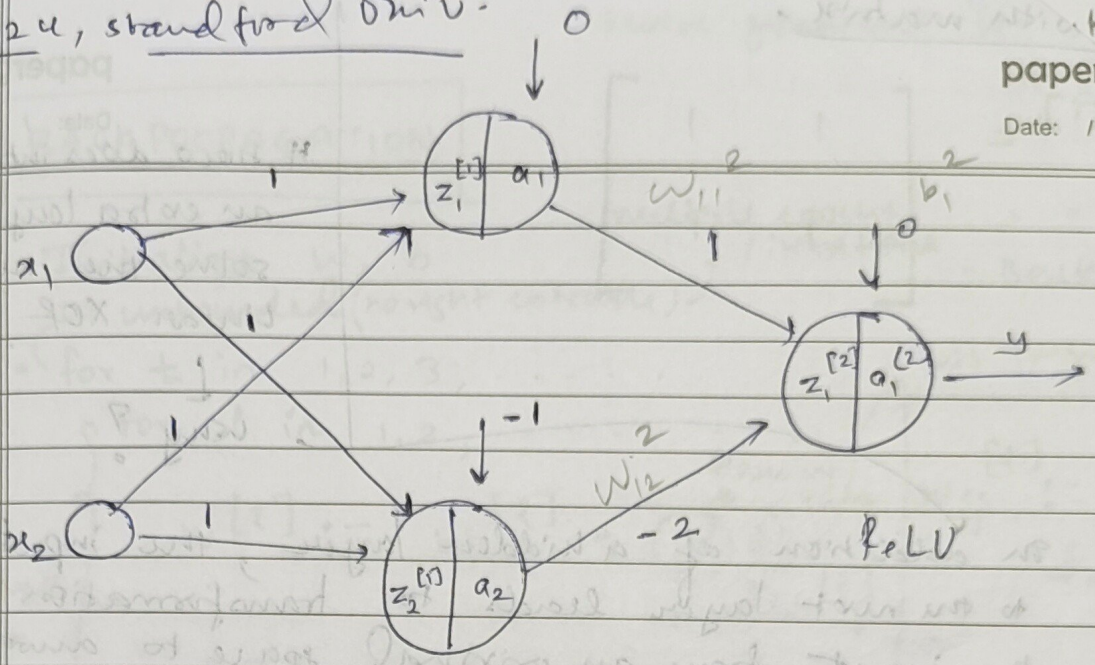
cs124, standford Univ.

Date: / / 25



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

$z_1^{[1]} = 1 \times 0 + 0 \times 1 + 0 = 0.$

$z_2^{[1]} = 1 \times 0 + 0 \times 1 + (-1) = -1$

$a_1(z_1^{[1]}) = ReLU(z_1^{[1]}) = 0$

$a_2(z_2^{[1]}) = ReLU(z_2^{[1]}) = 0.$

$z_1^{[2]} = w_{11}^2 \times a_1^1 + w_{12}^2 \times a_2^1 + b_1^2$

$= 1 \times 0 + (-2) \times 0 + 0$

$= 0$

$a_1^{(2)} = ReLU(z_1^{(2)}) = 0.$

$y = 0.$

x_1	x_2	y	$\hat{y}$
0	0	0	0
0	1	1	1
1	0	1	1
1	1	0	0

↳ handcrafting the initialization on the basis of domain knowledge.

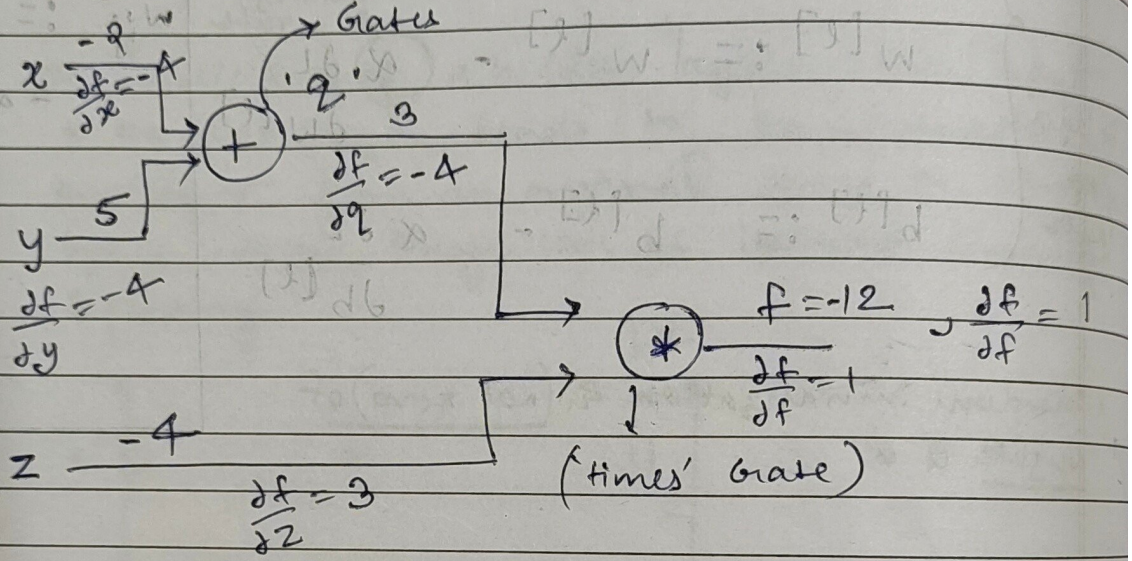
papergrid

Date: / /

Computational Graph.

~~Computation~~

* $f = (x + y)z$; let $q = x + y$
  $f = qz$



- How changes in the input leads to change in o/p
- $\frac{\partial f}{\partial x}$ → how x is influencing f ?
- $\frac{\partial f}{\partial x}$, $\frac{\partial f}{\partial y}$, $\frac{\partial f}{\partial z}$?

$$\frac{\partial q}{\partial x} = \frac{\partial (x+y)}{\partial x} = 1 \quad , \quad \frac{\partial q}{\partial y} = 1$$

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

Chain rule.

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

$$= \frac{\partial f}{\partial x} \cdot 4 \times 1$$

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

$$= -4 \times 1$$

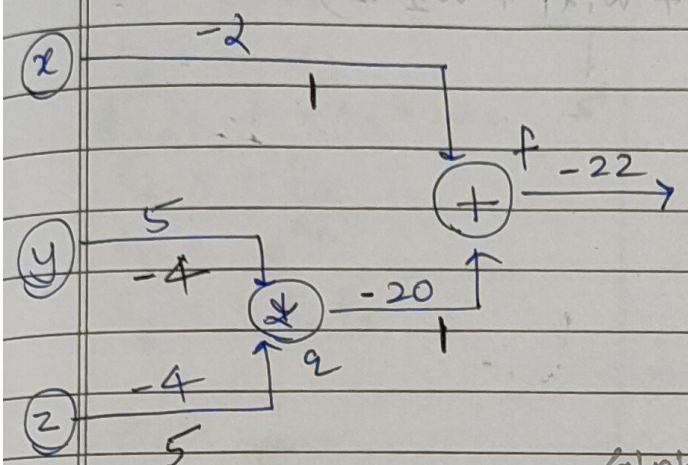
$$= -4$$

Define Additional functions. → {q = yz ; f = x + q}

Design Computational Graph

papergrid

Date: / /



Define forward pass.

q = 5x - 4 = -20

f = x + q = ~~-20~~ -2 + (-20)

= -22

Perform Backward pass → Global (gradient flowing from the final output)

{ $\frac{df}{dq} = \frac{d(x+q)}{dq} = 1$, $\frac{df}{dx} = 1$ } ~~# local~~ ← ~~Global~~.

{ $\frac{dq}{dy} = \frac{d(yz)}{dy} = z = -4$, $\frac{dq}{dz} = \frac{d(yz)}{dz} = y = 5$ } ~~# local~~.

Now, $\frac{df}{dy} = \frac{dq}{dy} \times \frac{df}{dq} = z \times 1 = z = -4$

$\frac{df}{dz} = \frac{dq}{dz} \times \frac{df}{dq} = y \times 1 = y = 5$

∴ Overall influence of x on f is 1, y on f is -4 & z on f is 5.

Give only 2 inputs to each gate

papergrid

Date: / /

$f_1 = w_0 x_0$

$f_2 = w_1 x_1$

$f_3 = f_1 + f_2 + w_2$

$f_4 = -f_3$

$f_5 = e^{f_4}$

$f_6 = 1 + f_5$

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

$f = f_7$

$f_1 = w_0 x_0$

$f_2 = w_1 x_1$

$f_3 = f_1 + f_2$

$f_4 = f_3 + w_2$

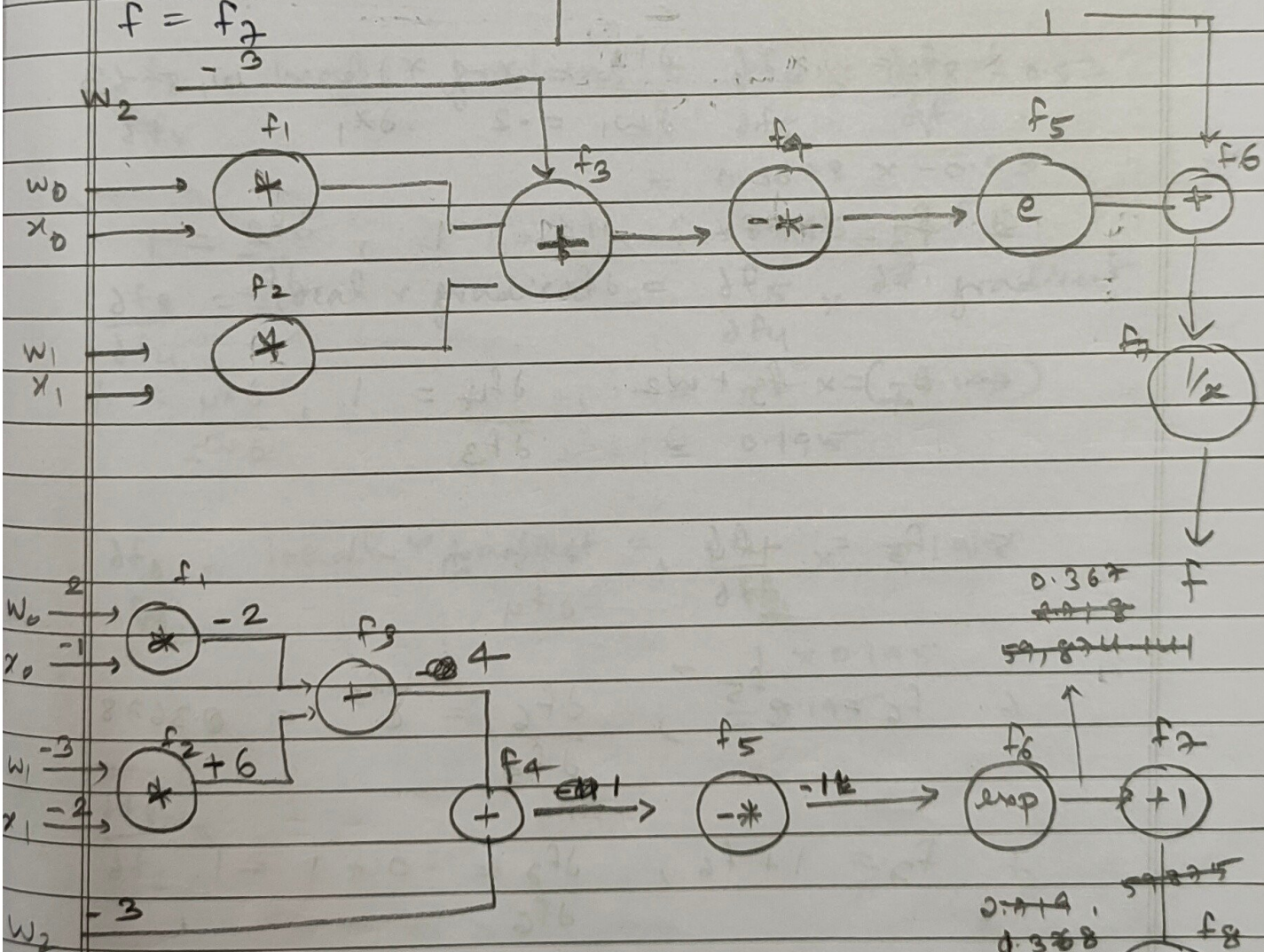
$f_5 = -f_4$

$f_6 = e^{f_5}$

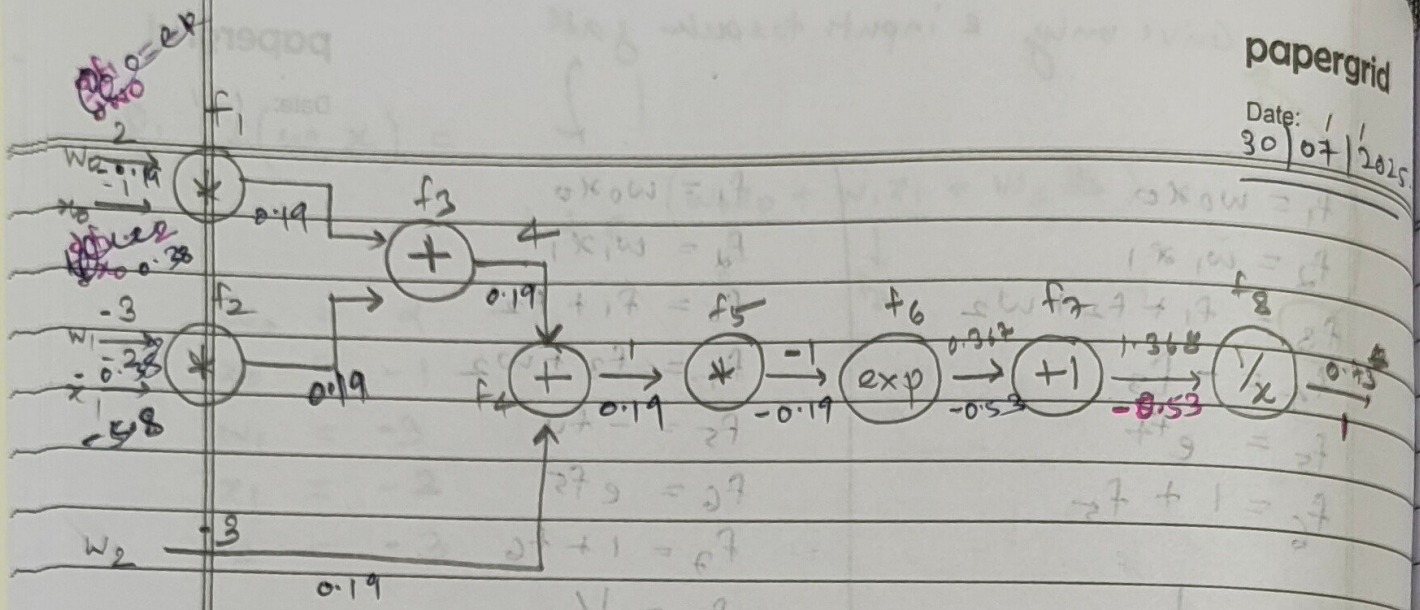
$f_7 = 1 + f_6$

$f_8 = 1/f_7$

~~$f = f_8$~~



Date: 30/07/2025



1. $f_1 = w_0 x_0$; $\frac{df_1}{dw_0} = x_0$, $\frac{df_1}{dx_0} = w_0 = 2$

2. $f_2 = w_1 x_1$; $\frac{df_2}{dw_1} = x_1$, $\frac{df_2}{dx_1} = w_1 = -3$

3. $f_3 = f_1 + f_2$; $\frac{df_3}{df_1} = 1$, $\frac{df_3}{df_2} = 1$

4. $f_4 = f_3 + w_2$; $\frac{df_4}{df_3} = 1$, $\frac{df_4}{dw_2} = 1$

5. $f_5 = -f_4$, $\frac{df_5}{df_4} = -1$

6. $f_6 = e^{f_5}$, $\frac{df_6}{df_5} = e^{f_5} = 0.3678$

7. $f_7 = 1 + f_6$, $\frac{df_7}{df_6} = 0 + 1 = 1$

8. $f_8 = 1/f_7$, $\frac{df_8}{df_7} = \frac{d(f_7^{-1})}{d(f_7)} = \frac{-1}{(f_7)^2} = -0.53$

Back prop:

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

$$= -\frac{1}{(f_7)^2} \times 1 = -0.53$$

$$\frac{df_8}{df_6} = \text{local} \times \text{gradient} = \frac{df_7}{df_6} \times \frac{df_8}{df_7}$$

$$= 1 \times (-0.53)$$

$$= -0.53$$

$$\frac{df_8}{df_5} = \text{local} \times \text{gradient} = \frac{df_6}{df_5} \times \frac{df_8}{df} - 0.53$$

$$= 0.3678 \times -0.53$$

$$= -0.195$$

$$\frac{df_8}{df_4} = \text{local} \times \text{gradient} = \frac{df_5}{df_4} \times \text{gradient}$$

$$= -1 \times (-0.195)$$

$$= 0.195$$

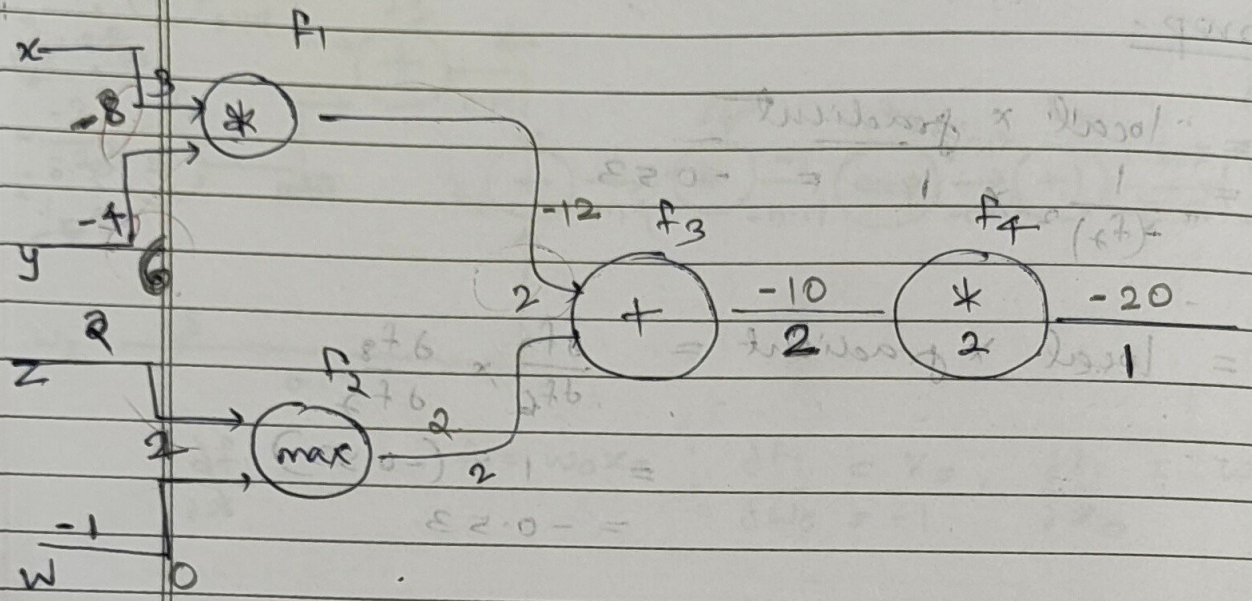
$$\frac{df_8}{df_2} = \text{local} \times \text{gradient} = \frac{df_4}{df_2} \times 0.195$$

$$= 1 \times 0.195$$

$$= 0.195$$

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

#



$f_1 = x * y$

$\frac{\partial f_1}{\partial x} = y, \quad \frac{\partial f_1}{\partial y} = x$

$f_2 = \max(z, w)$

$f_3 = f_1 + f_2$

$f_4 = 2 \times f_3$

$\frac{\partial f_2}{\partial z} = \begin{cases} 1 & \text{if } z > w \\ 0 & \text{if } z < w \\ \text{undefined} & \text{if } z = w. \end{cases}$

lead:

$\frac{\partial f_2}{\partial z} = \frac{\partial(\max(z,w))}{\partial z}; \; z > w$

then due to partial derivative $w \to$ const.

$= \frac{\partial(z)}{\partial z} = 1$

$\frac{\partial f_2}{\partial z}\Big|_{z<w} = \frac{\partial(w)}{\partial z} = 0$

$\frac{\partial f_4}{\partial f_3} \neq \frac{\partial f_3}{\partial f_2}$ ✗

$\frac{\partial f_3}{\partial f_1} = 1$

$\frac{\partial f_3}{\partial f_2} = 1$

$\frac{\partial f_4}{\partial f_3} = 2$

$\frac{\partial f_2}{\partial w} = \begin{cases} 1 & \text{if } w > z \\ 0 & \text{if } w < z \\ \text{undefined} & \text{if } w = z. \end{cases}$

Date: / /

$\frac{\partial f_4}{\partial f_3} = 2$, $\frac{\partial f_4}{\partial f_2} = \frac{\partial f_3}{\partial f_2} \times \frac{\partial f_4}{\partial f_3} = 1 \times 2 = 2$

$\frac{\partial f_4}{\partial f_1} = \frac{\partial f_3}{\partial f_1} \times 2 = 2$

$\frac{\partial f_4}{\partial f_2} = 2$

$\frac{\partial f_4}{\partial 2} = \frac{\partial f_2}{\partial 2} \times 2 = 2 \times 1 = 2$

remember.

31/07/2021.

Note -

(1.) (+) gate acts as a gradient distributor., the gradient gets distributed equally to all the inputs. ✓

(2.) (max) gate, in this gate the gradient flows towards the path where the input values are higher. ✓

In previous eg:, f_2 will have no impact or change of ∞ input. ✓

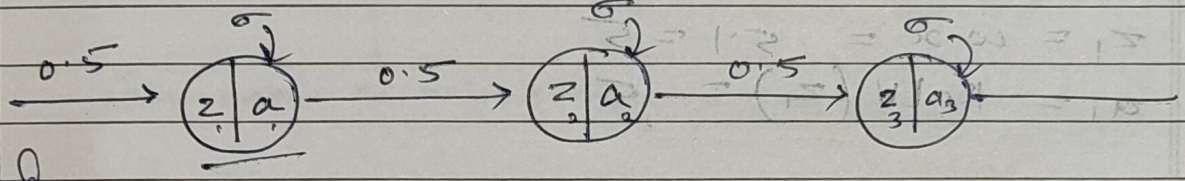
~~routes the gradient to the hi~~

(3.) (*) gate → swaps the input activations & multiply by global gradient. (applicable with two inputs only) **

(times or multiply gate)

- ReLU → as it prevents saturation of +ve inputs, used to fix the Vanishing Grad problem.
- Opposite of Vanishing → Exploding gradients → caused by ReLU.
- To fix exploding gradients, caused by ReLU, 'clipping' technique is used. → for gradients.

Example.



Forward flow:

$z_1 = 1 \times 0.5 = 0.5$

$a_1 = \sigma(z_1) = \frac{1}{1+e^{-0.5}} = \frac{1}{1+0.606} = \frac{1}{1.606}$

$= 0.622$

$z_2 = 0.5 \times 0.622 = 0.311$

$a_2 = \sigma(z_2) = \frac{1}{1+e^{-0.311}} = \frac{1}{1+0.732} = \frac{1}{1.732}$

$= 0.577$

$z_3 = 0.577 \times 0.5 = 0.288$

$a_3 = \frac{1}{1+e^{-0.288}} = \frac{1}{1.749} = 0.571$

$L = \frac{1}{2}(a_3 - y)^2 \quad \therefore \frac{\partial L}{\partial a_3} = a_3 - y = 0.572 - 1$

$= -0.428.$

$\frac{\partial a_1}{\partial x}$ =

22nd ICDCIT 2026 ← Abstract submission (Aug 30)
CODS 2025 ← (Sept 15)

papergrid
Date: / /

During the test time, the entire architecture is kept as it is, the entire data is shown to the neurons. To account for the keep probability at the training time, a scaling factor is associated at each neuron during testing.

~~* scaling factor is keep prob. multiplied to the activation~~

→ Standard Dropout. → scaling factor is keep prob. * to the activation / weights.

no scaling at test-time.

* Inverted Dropout. → to maintain consistent activation magnitudes

entire net.

testing time.

↳ no dropout.

$w_1 = 2$

p = 0.5

training time

1 → output = 2
2 → (X) dropout → output = 0
3 → (X) output = 0
4 → output = 2

Total output = 2 + 2 = (4)

multiply by 1/p

$\therefore 4 \times \frac{1}{0.5} = 8$

Sample 1 → o/p → 2
Sample 2 → o/p → 2
Sample 3 → o/p → 2
Sample 4 → o/p → 2

Total o/p = 8. ✓

Scale the activation during training by multiplying with (1/0.5) or 1/p

high pixel → feature

values represent how the feature is detected in the original image.

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

$*$ (convolution operator)

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

Filter or kernel (3×3) → learn this parameter?

$=$

-5	-4	0	8
-10	-2	2	3
0	-2	-4	-7
-3	-2	-3	+6

⇓

{ feature map } = Activation map

input 6×6 image

Element wise multiplication & summation

→ Take the filter & superimpose on the input.

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

$= -5$

→ slide the filter by one

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

$= -2 - 4 + 5$

$= -2 + 1$

$= -4$

3×3 14/08/2025

Example.

0	0	0	0	0	0	0
0	2	3	7	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	4	2	1	8	3	0
0	0	0	0	0	0	0

$n = 5 \times 5$ $\quad 7 \times 7$

$p = 1$

$s = 2$

$* \begin{bmatrix} 1 & -1 \\ 0 & -1 \end{bmatrix}_{2\times 2} = \begin{bmatrix} -2 & -7 & -6 \\ -9 & -11 & -7 \\ -11 & 4 & -3 \end{bmatrix}_{3\times 3}$

$\left\lfloor \frac{5 + 2\times 1 - 2 + 1}{2} \right\rfloor = 3$

$\left\lfloor \frac{n + 2p - f + 1}{s} \right\rfloor * \left\lfloor \frac{n + 2p - f + 1}{s} \right\rfloor$

3×3

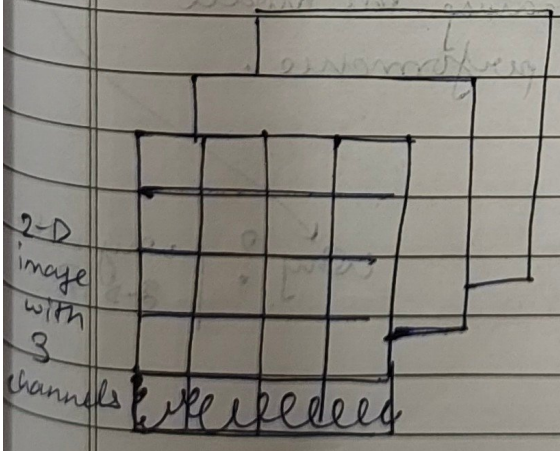
***General formula :- $\left\lfloor \frac{n + 2p - f + 1}{s} \right\rfloor$

$\frac{7 + 4 - 2}{2} = \frac{9}{2}$

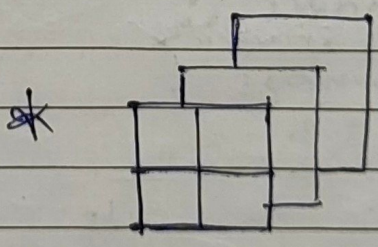
floor

$\left\lfloor \frac{n + 2p - f}{s} + 1 \right\rfloor$

Convolutions on 3D Volume. → eg: RGB.



2-D image with 3 channels



$2 \times 2 \times 3$

$4 \times 4 \times 3$

height, width, no. of channels

- filters for each channel stacked behind each other.

q) 10 filters } in one layer of Conv. Net
3 × 3 × 3 }

How many parameters does this layers have?

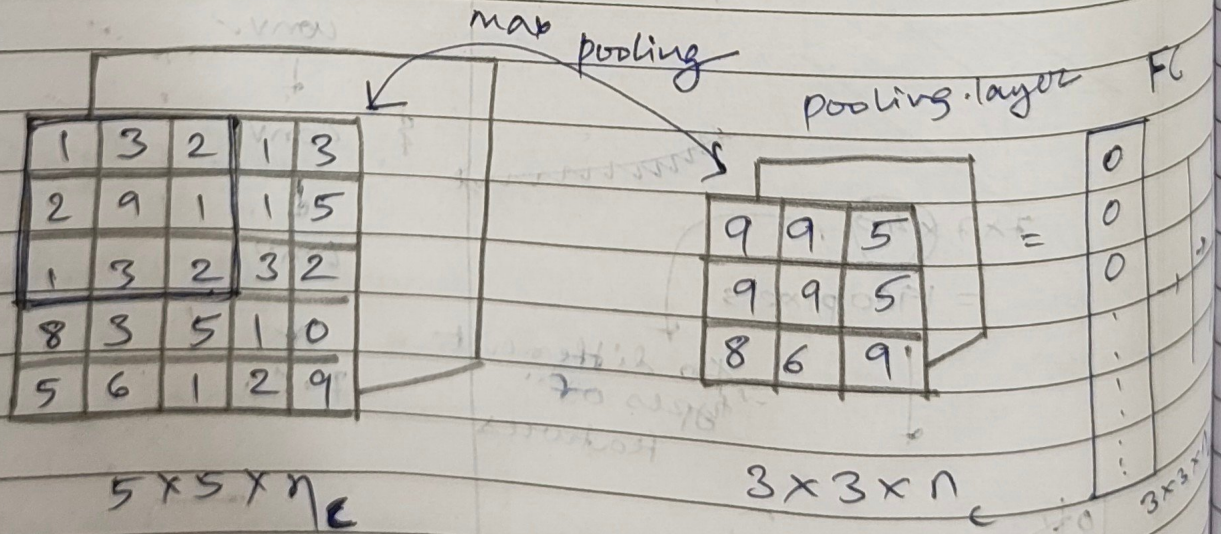
↳ weights & biases.

↓ 270 ↓ 10

1 filter ⟶ 27 weights + 1 bias.

10 filter ⟶ 270 w + 10 b = 280. ✓

* 1. CONV layer.
2. Pooling layer ⟶ Max Pooling / ~~Avg. Pooling~~
3. Fully connected layer (FC)



CONV layer ⟶ pooling layer ⟶ CONV layer

Character level language model.

"Hello World"

$\tau = 2$ (two time steps)

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

input dim $= 2 \Rightarrow x_t \in \mathbb{R}^2$

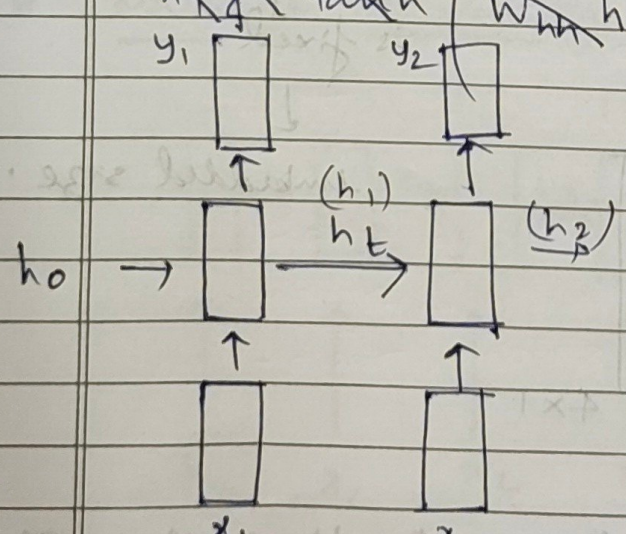
hidden state dim $= 3 \Rightarrow h_t \in \mathbb{R}^3$

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

$y_t \in \mathbb{R}^2$

$W_{hy} = \begin{bmatrix} 1.0 & -1.0 & 0.5 \\ 0.5 & 0.5 & -0.5 \end{bmatrix}_{2\times3}$

$h_t = g(W_{hh} h + W_{xh} x)$



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

$= \begin{bmatrix} 0.1 & 0.4 & 0.0 \\ -0.2 & 0.3 & 0.2 \\ 0.05 & -0.1 & 0.2 \end{bmatrix}_{3\times3} \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}_{3\times1} + \begin{bmatrix} 0.5 & -0.3 \\ 0.8 & 0.2 \\ 0.1 & 0.4 \end{bmatrix}_{3\times2} \begin{bmatrix} 1 \\ 2 \end{bmatrix}_{2\times1}$

① $h_1 = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} + \begin{bmatrix} -0.1 \\ 1.2 \\ 0.9 \end{bmatrix} = \begin{bmatrix} -0.1 \\ 1.2 \\ 0.9 \end{bmatrix}_{3\times1}$

$\frac{-0.1 - 1.2 + 0.4r}{-1.3 + 0.4r}$

papergrid

Date: / /

$$y_1 = W_{hy} \times h_1 = \begin{bmatrix} 1.0 & -1.0 & 0.5 \\ 0.5 & 0.5 & -0.5 \end{bmatrix}_{3\times3} \begin{bmatrix} -0.1 \\ 1.2 \\ 0.9 \end{bmatrix}_{3\times1}$$

$$= \begin{bmatrix} -085 \\ 0.1 \end{bmatrix}_{2\times1}$$

$$h_2 = \begin{bmatrix} 0.1 & 0.4 & 0.0 \\ -0.2 & 0.3 & 0.2 \\ 0.0r & -0.1 & 0.2 \end{bmatrix} \begin{bmatrix} -0.1 \\ 1.2 \\ 0.9 \end{bmatrix} + \begin{bmatrix} 0.5 & -0.3 \\ 0.8 & 0.2 \\ 0.1 & 0.4 \end{bmatrix} \begin{bmatrix} -1 \\ 1 \end{bmatrix}$$

$$-0.01 + 0.48 + 0$$

$$+0.02$$

$$h_1 = \tanh\left(\begin{bmatrix} -0.1 \\ 1.2 \\ 0.9 \end{bmatrix}\right)$$

$$g(z) = \frac{e^z - e^{-z}}{e^z + e^{-z}}$$

$$g(-0.1) = \frac{1.105 - 0.904}{1.105 + 0.904}$$

$$= \frac{0.201}{2.}$$

$$h_1 = \begin{bmatrix} -0.1 \\ 0.8 \\ 0.7 \end{bmatrix}$$

$$g(1.2) = \frac{3.32 - 0.301}{3.32 + 0.301}$$

$$g(0.9) = \frac{2.5 - 0.406}{2.906} = \frac{3.019}{3.621}$$

$$= 0.834$$

$$= 0.72$$

$$y_1 = W_{hy} \times h_1 = \begin{bmatrix} 1.0 & -1.0 & 0.5 \\ 0.5 & 0.5 & -0.5 \end{bmatrix} \begin{bmatrix} -0.1 \\ 0.8 \\ 0.7 \end{bmatrix}$$

$$= \begin{bmatrix} -0.1 + (-0.8) + (0.35) \\ -0.05 + 0.4 - 0.35 \end{bmatrix}$$

$$= \begin{bmatrix} -0.35 \\ 0 \end{bmatrix} \quad \begin{bmatrix} -0.55 \\ 0.008 \end{bmatrix}$$

$$= \begin{bmatrix} -0.1 - 0.8 + 0.35 \\ -0.05 + 0.40 - 0.35 \end{bmatrix} = \begin{bmatrix} -0.55 \\ 0.008 \end{bmatrix}$$

$$h_2 = \begin{bmatrix} 0.1 & 0.4 & 0.0 \\ -0.2 & 0.3 & 0.2 \\ 0.05 & -0.1 & 0.2 \end{bmatrix} \begin{bmatrix} -0.099 \\ 0.83 \\ 0.716 \end{bmatrix} + \begin{bmatrix} 0.5 & -0.3 \\ 0.8 & 0.2 \\ 0.1 & 0.4 \end{bmatrix} \begin{bmatrix} -1 \\ + \\ - \end{bmatrix}$$

$$= \begin{bmatrix} -0.01 + 0.32 + 0 \\ -0.02 + 0.24 + 0.14 \\ -0.005 - 0.08 + 0.14 \end{bmatrix} + \begin{bmatrix} -0.5 - 0.3 \\ -0.8 + 0.2 \\ -0.1 + 0.4 \end{bmatrix}$$

$$= \begin{bmatrix} 0.31 \\ 0.36 \\ 0.056 \end{bmatrix} + \begin{bmatrix} -0.8 \\ -0.6 \\ 0.3 \end{bmatrix} = \begin{matrix} -0.5 \\ -0.24 \\ 0.356 \checkmark \end{matrix}$$

Date: / /

$$-0.0099 + 0.332 + 0$$
$$0.00198 + 0.249 + 0.1432$$
$$-0.0049 - 0.083 + 0.1432$$

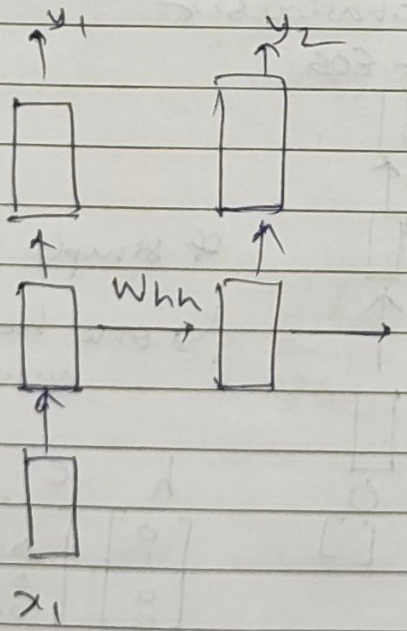
$$\begin{bmatrix} 0.3221 \\ 0.3941 \\ 0.0553 \end{bmatrix} + \begin{bmatrix} -0.8 \\ -0.6 \\ 0.3 \end{bmatrix}$$

$$h_2 = \begin{bmatrix} -0.4778 \\ -0.2059 \\ 0.3553 \end{bmatrix}$$

$$y_2 = \begin{bmatrix} -0.0886 \\ -0.4844 \end{bmatrix}$$

Weights used are the same, therefore parameters are shared across the time stamps in RNN.

How does this help?



$$\frac{\partial L}{\partial W_{hh}}.$$

Example:

$x_t = [0.5, -0.1]^T$

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

$c_t = [0.2, -0.2]^T$

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

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

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

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

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

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

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

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

$\{ h_t = ?, \; c_t = ? \}$

$h_t = o_t \odot \tanh(c_t)$ → ~~dot products~~ Element wise multiplication.

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

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

papergrid3

Date: 1/0.28

$i_t \rightarrow x_t \rightarrow c_t \rightarrow f_t \rightarrow o_t \rightarrow h_t$

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

$$= \sigma\left(\begin{bmatrix} 0.1 & 0.2 \\ -0.2 & 0.05 \end{bmatrix}_{2\times 2} \begin{bmatrix} 0.0 \\ 0.1 \end{bmatrix}_{2\times 1} + \begin{bmatrix} 0.5 & -0.3 \\ 0.4 & 0.1 \end{bmatrix} \begin{bmatrix} 0.5 \\ -0.1 \end{bmatrix}\right)$$

$$= \sigma\left(\begin{bmatrix} -0.02 \\ 0.005 \end{bmatrix} + \begin{bmatrix} 0.28 \\ 0.49 \end{bmatrix}\right) = \sigma\begin{bmatrix} 0.30 \\ 0.495 \end{bmatrix}$$

$$i_t = \begin{bmatrix} 0.57 \\ 0.59 \end{bmatrix}$$

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

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

$$= \tanh\left(\begin{bmatrix} 0.01 \\ 0.005 \end{bmatrix} + \begin{bmatrix} -0.29 \\ 0.49 \end{bmatrix}\right) = \begin{bmatrix} -0.28 \\ 0.495 \end{bmatrix}$$

$$= \begin{bmatrix} -0.272 \\ 0.13 \end{bmatrix}$$

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

$f_t = \sigma\left(W_{hf} h_{t-1} + W_{xf} x_t\right)$

$= \sigma\left(\begin{bmatrix} 0.05 & -0.1 \\ 0.2 & 0.1 \end{bmatrix} \begin{bmatrix} 0.6 \\ 0.1 \end{bmatrix} + \begin{bmatrix} -0.4 & 0.2 \\ 0.3 & 0.3 \end{bmatrix} \begin{bmatrix} 0.5 \\ -0.1 \end{bmatrix} \right)$

$= \sigma\left(\begin{bmatrix} -0.01 \\ 0.01 \end{bmatrix} + \begin{bmatrix} -0.22 \\ 0.12 \end{bmatrix} \right) = \sigma\begin{pmatrix} -0.23 \\ 0.13 \end{pmatrix}$

$= \begin{bmatrix} 0.44 \\ 0.53 \end{bmatrix}$

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

$= \tanh\left(\begin{bmatrix} 0.2 & 0.1 \\ -0.1 & 0.05 \end{bmatrix} \begin{bmatrix} 0.6 \\ 0.1 \end{bmatrix} + \begin{bmatrix} -0.5 & 0.4 \\ 0.2 & -0.3 \end{bmatrix} \begin{bmatrix} 0.5 \\ -0.1 \end{bmatrix} \right)$

$= \begin{bmatrix} -0.27 \\ 0.13 \end{bmatrix}$

$o_t = \begin{bmatrix} 0.53 \\ 0.46 \end{bmatrix}$

output of gate is passed using sigmoid → Therefore b/w 0-1 ;

$$C_t = f_t \odot C_{t-1} + i_t \odot g_t$$

$$= \begin{bmatrix} 0.44 \\ 0.53 \end{bmatrix} \odot \begin{bmatrix} 0.2 \\ -0.2 \end{bmatrix} + \begin{bmatrix} 0.57 \\ 0.55 \end{bmatrix} \odot \begin{bmatrix} -0.27 \\ 0.13 \end{bmatrix}$$

$$C_t = \begin{bmatrix} 0.088 \\ -0.106 \end{bmatrix} + \begin{bmatrix} -0.1539 \\ 0.0715 \end{bmatrix} = \begin{bmatrix} -0.0659 \\ -0.0345 \end{bmatrix}$$

$$h_t = o_t \odot \tanh(C_t)$$

$$= \begin{bmatrix} 0.53 \\ 0.46 \end{bmatrix} \odot \tanh \begin{bmatrix} -0.0659 \\ -0.0345 \end{bmatrix}$$

$$= \begin{bmatrix} 0.53 \\ 0.46 \end{bmatrix} \odot \begin{bmatrix} -0.0658 \\ -0.0344 \end{bmatrix}$$

$$h_t = \begin{bmatrix} -0.0348 \\ -0.0158 \end{bmatrix} = \begin{bmatrix} -0.03 \\ -0.01 \end{bmatrix}$$

Interview tips. 1) Understand the Dataset properly
2) • Seurat Bioinfo tools. papergrid
• Generative models (Data simulation)

Date: / /

$x \in \mathbb{R}^2$, $x = \begin{bmatrix} 2 \\ 0 \end{bmatrix}$ 30/09/2025

$h \in \mathbb{R}^1$; what should be size of W_e & W_d.

context vector

- Denoising Autoencoders
- Attention

$W_e \rightarrow 1 \times 2$, $W_d \rightarrow 2 \times 1 = \begin{bmatrix} 1.0 \\ 0.5 \end{bmatrix}_{2\times1}$

$= \begin{bmatrix} 0.5 & -1.0 \end{bmatrix}_{1\times2}$

$\text{MSE loss} = \frac{1}{2}\left\| \hat{x} - x \right\|_2^2$

Autoencoder - ~~encoder~~ $h = xW_e = \begin{bmatrix} 2 \\ 0 \end{bmatrix}_{2\times1} \begin{bmatrix} 0.5 & -1.0 \end{bmatrix}_{1\times2}$

$h = 1$

decoder

$x = o = h \cdot W_d = 1 \times \begin{bmatrix} 1.0 \\ 0.5 \end{bmatrix}$

loss

$L = \sqrt{(-1)^2 + |0.25|} = \sqrt{0.25} = \frac{1}{2} = 0.5$

$L = (1-2)^2 + (0.5-0)^2 = 1.25$

$L = \frac{1}{2}\left[(1-2)^2 + (0.5-0)^2\right] = 0.625$

[1, 0, 1] [1 0 1

rgrid

25

2) $h_1 = [1, 0, 1]$ scoring function = DO

$h_2 = [0, 1, 1]$

$h_3 = [1, 1, 0]$ $c_t = ?$

$s_{t-1} = [1, 0, 1]$

$$\alpha_{t,j} = \frac{\exp(\text{score}(s_{t-1}, h_j))}{\sum_{j=1}^{T} \exp(\text{score}(s_{t-1}, h_j))}$$

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

i) $$\alpha_{t,1} = \frac{\exp([1,0,1] \cdot [1,0,1])}{\exp([1,0,1] \cdot [1,0,1]) + \exp([1,0,1] \cdot [0,1,1]) + \exp([1,0,1] \cdot [1,1,0])}$$

$$= \frac{\exp(2)}{\exp(2) + \exp(1) + \exp(1)} = \frac{5.43}{5.43 + 2.718 + 2.718}$$

$$\alpha_{t,1} = \frac{5.43}{10.86} = 0.5$$

$$= \frac{7.389}{7.389 + 2.71 + 2.71} = 0.576.$$

$\alpha_{t,2} = 0.2$

$\alpha_{t,3} = 0.2$

$$c_t = 0.576 \times [1,0,1] + 0.2 \times [0,1,1] + 0.2 \times [1,1,0]$$

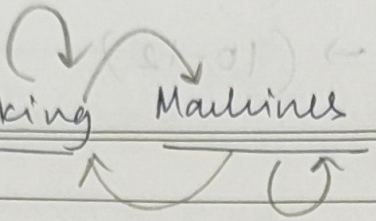
$$= [0.576, 0, 0.576] + [0.2, 0.2, 0.2] + [0.2, 0.2, 0]$$

$$c_t = [0.776, 0.4, 0.776]^T$$

* $[c_t$ closer to $h_1]$ → sets aligned to one of the hidden states.

focuses on this part of input to generate next word (below).

Date: 09/10/2025



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 = ~~Machines~~ ~~Thinking~~ Thinking

② q_1 = Machines
k_2 = Machines

i/p ↘

Playing outside ⇒ o/p ↓ z_1 & z_2

Playing.

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

$k_1 = [0.31 \quad 0.84 \quad 0.983 \quad 0.57]^T$

$v_1 = [0.36 \quad 0.83 \quad 0.1 \quad 0.35]^T$

outside

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

$k_2 = [0.45 \quad 0.94 \quad 0.73 \quad 0.58]^T$

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

playing.

score $q_1 \times k_1$ $q_1 \times k_2$

$\downarrow$

$[0.212\ 0.04\ 0.63\ 0.36]^T$

score 0.91121 0.8017.

divide by $\sqrt{d_k}$ $\frac{0.91121}{\sqrt{4}} = 0.456$ $\frac{0.8012}{\sqrt{4}} = 0.401$

softmax $\frac{e^{-0.456}}{e^{-0.456} + e^{-0.401}}$ $\frac{e^{-0.401}}{e^{-0.456} + e^{-0.401}}$

$= \frac{0.633}{0.633 + 0.669}$ $\frac{0.669}{0.633 + 0.669}$

$= \frac{0.633}{1.302}$ $\frac{0.669}{1.302}$

$= 0.486$ $= 0.5138$

$Z_1 =$ softmax $\times V_1$ + softmax $\times V_2$

$= 0.486 \times [0.36\ 0.83\ 0.1\ 0.35]^T$

$+ 0.5138 \times [0.31\ 0.36\ 0.19\ 0.72]^T$

$= [0.17496\ 0.40338\ 0.0486\ 0.1701]$

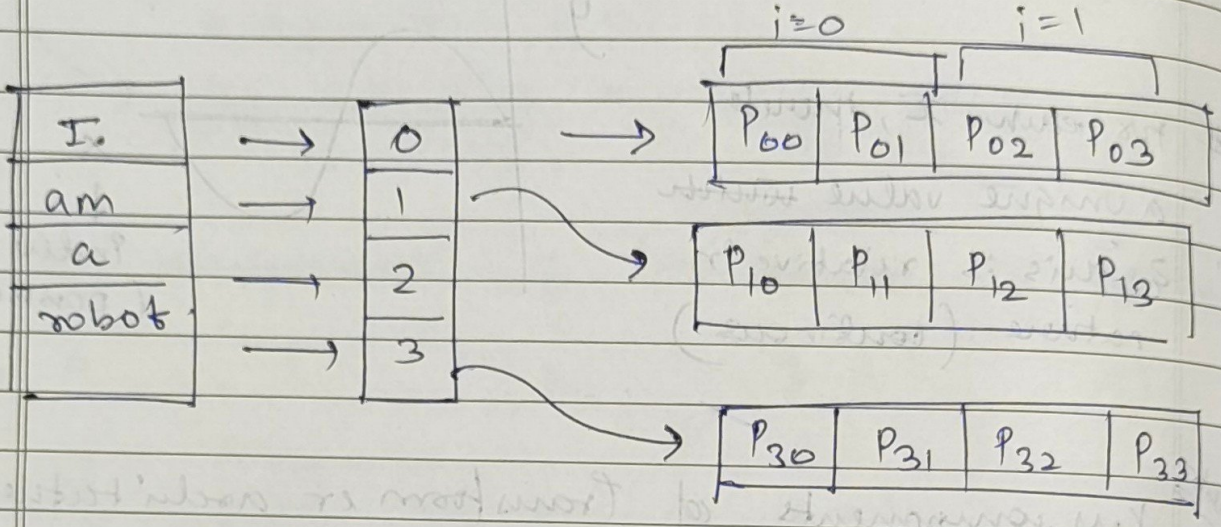
$+ [0.1592\ 0.1849\ 0.097\ 0.3699]$

P_{00} → the first index.
→ index within the word for each element

papergrid Date

Example.

I am a robot. $P_\epsilon : d = 4 \rightarrow \mathbb{R}$



$$PE_{pos,2i} = \sin\left(\frac{pos}{100^{(2i/d_{emb})}}\right)$$

For $i = 0$: (position) $1^0 = 0$.

$$PE_{0,0} = \sin\left(\frac{0}{100^{(2\times0/4)}}\right) = \sin(0^\circ) = 0.$$

$0,0$ ← dimension index.

$$PE_{0,1} = \cos(0) = \cos(0) = 1$$

$$PE_{0,2} = \sin\left(\frac{0}{100^{(2\times0/4)}}\right) = \sin(0) = 0$$

$$PE_{0,3} = \cos\left(\frac{0}{100^{(2i/d_{emb})}}\right) = \cos(0) = 1$$

For $i = 1$

$$P_{10} = \sin\left(\frac{1}{100(2/4)}\right) = \sin\left(\frac{1}{100^{1/2}}\right)$$

$$= \sin\left(\frac{0.1}{0.02}\right)$$

$$= 0.0017.$$

$$P_{11} = \cos\left(\frac{1}{100(1/2)}\right) = \cos(0.1) = 0.99$$

$$P_{12} = \sin\left(\frac{1}{100(2/4)}\right) = \sin(0.1) = 0.0017.$$

$$P_{13} = \cos(0.1) = 0.99.$$

For $i = 3$.

$$P_{30} = \sin\left(\frac{3}{100(3/4)}\right) = \sin\left(\frac{3}{(100)^{1/2}}\right)$$

$$= \sin\left(\frac{3}{\sqrt[2]{100}}\right)$$

$$= \sin\left(\frac{3}{1000}\right)$$

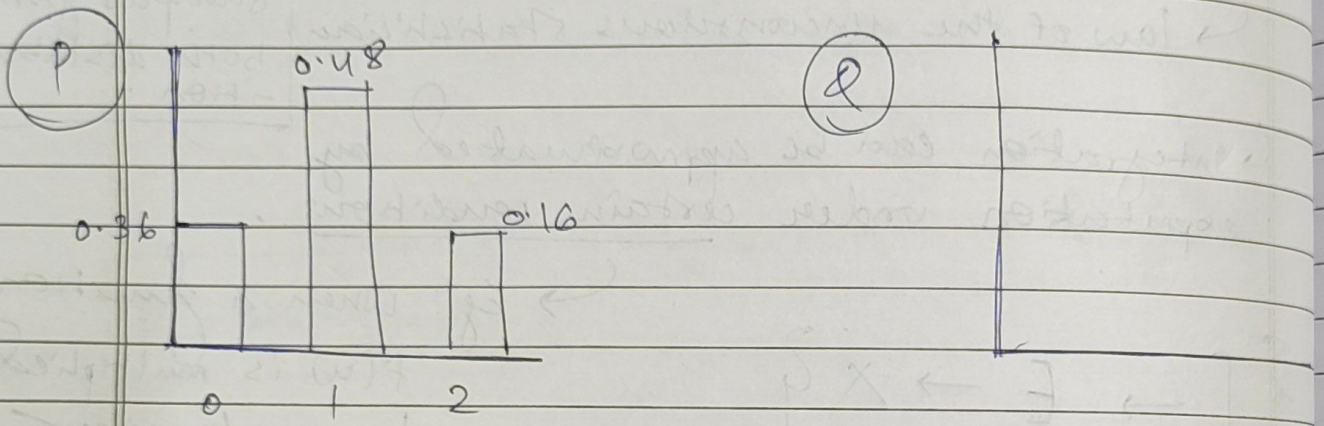
$$= \sin(0.003)$$

$$= 5.23 \times 10^{-5}$$

$$\underset{\theta \in M}{\operatorname{argmin}} \; \underset{X \sim P_X}{\mathbb{E}} \left[-\log \left(P_\theta(x) \right) \right]$$

{ Maximum likelihood Estimation. }

Example -



Binomial distribution

Distribution X	0	1	2
P(X)	9/25	12/25	4/25
Q(X)	1/3	1/3	1/3

$$D_{KL}(P \,||\, Q) = \sum_{x \in X} P(X=x) \ln \frac{P(X=x)}{Q(X=x)}$$

$$= \frac{9}{25} \ln\left(\frac{9/25}{1/3}\right) + \frac{12}{25} \ln\left(\frac{12/25}{1/3}\right) + \frac{4}{25} \ln\left(\frac{4/25}{1/3}\right)$$

$\approx 0.0852.$

$$D_{KL}(Q \| P) = \frac{9}{25}\cdot\frac{1}{3}\ln\left(\frac{1/3}{9/25}\right) + \frac{1}{3}\ln\left(\frac{1/3}{12/25}\right) + \frac{1}{3}\ln\left(\frac{1/3}{4/25}\right)$$

$$= \frac{1}{3}\ln\left(\frac{1}{3}\times\frac{25}{\cdots}\right)$$

$$= \frac{1}{3}\times 0.076 + \frac{1}{3}\times 0.36 + \frac{1}{3}\times 0.733$$

$$D_{KL}(Q \| P) = 0.09$$

* Two kinds of Generative Models.

i) Fully visible models. → considering each pixel value as an observation.

ii) latent variable models.