3rd March 2025:

Sequences:

MLP, CNN } Fixed length data

Language Translation. Length of i/p & o/p not fixed.

→ Images of Varying resolution.

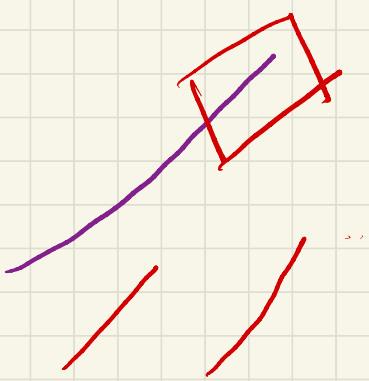
medical records.

Varying length i/p : → fixed length problem.

"fixing the len sequence"

Collection of items in an order : Sequences.

Day	Temp
1	30
2	32
3	33
4	31
5	?

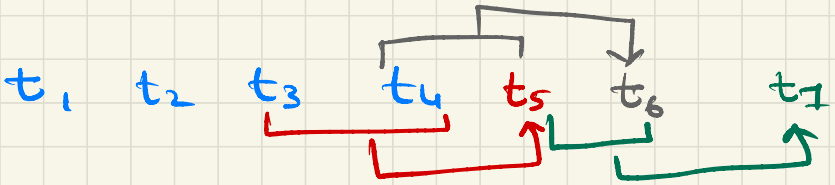


how much to go back?

taking only the past 2 days.

$$t_5 = b + w_1 t_4 + w_2 t_3$$

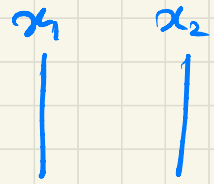
$t_6 = b + w_1 t_5 + w_2 t_4$



Auto Regressive models.

Auto → Self

Regressive → looks back in time.



height	weight
h_1	s_1
h_2	s_2
h_3	⋮
⋮	
h_m	s_m

Same

height	weight
h_1	s_1
h_m	s_m
h_2	s_2
h_3	s_3
⋮	⋮

t_1 t_2 t_3 t_4 t_5

t_2 t_5 t_3 t_4 t_1 ✗

t_0 t_1 t_2 t_3 t_4 t_5 t_6

X			Y
t_0	t_1	t_2	t_3
t_1	t_2	t_3	t_4
t_2	t_3	t_4	t_5
t_3	t_4	t_5	t_6

fix "p"

$$AR(p): \quad x_t = b + \sum_{i=1}^{p} w_i x_{t-i}$$

$p = 1$:

AR (1) :

$$x_t = b + w_1 x_{t-1}$$

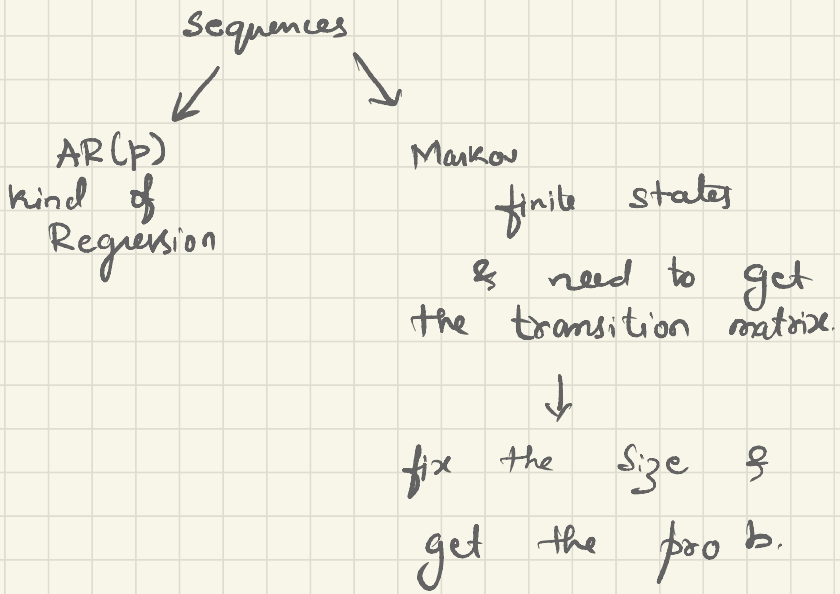
$$x_2 = b + w_1 x_1$$

$$x_3 = b + w_1 x_2$$

Markov:

o/p: next State

how many next states ?



Text to Numbers ?

I	am	enjoying	my	studies.
0	1	2	3	4

(821)

I	love	my	country.
0	1	2	3

(242)

Oxford English Dict.

1 world

2

3

⋮

242 : country.

enjoy : 821

Coffee is great

Car battery is dead, oh great

each word some number based on the index of constructed dictionary

I am from Mysore
My name is Raghava
I like eating Dosa.

- Tokenisation:

Vocabulary

6th Mar 2025:

fixed length i/p → fixed length o/p

→ Variable length i/p → fixed length o/p

Named Entity Representation: NER

XX Harry potter won the Triwizard Cup. X X

1 1 0 0 1 0

x_1 x_2 x_3 x_4 x_5 x_6

x_t : t^{th} feature

$x_{i,j}$: j^{th} feature in i^{th} example.

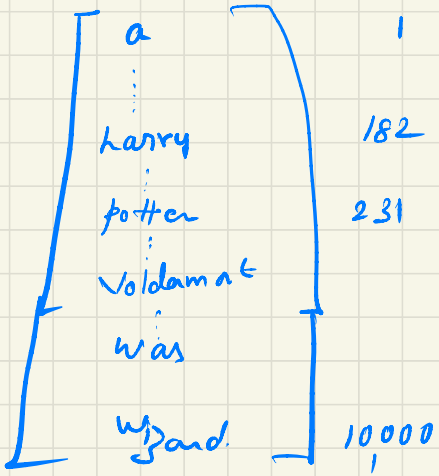
In this specific case of NER.

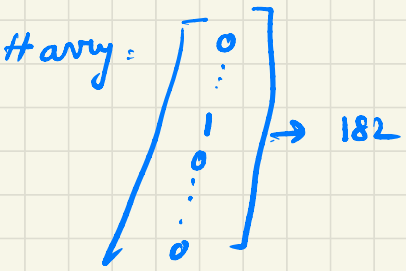
$y_1, y_2 \cdots, y_6$

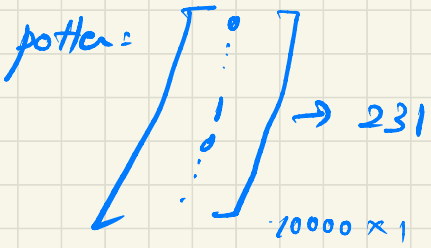
$y_{i,j}$: j^{th} output of i^{th} feature

Construct Vocabulary:

num of unique tokens (Size of vocabulary) is 10,000







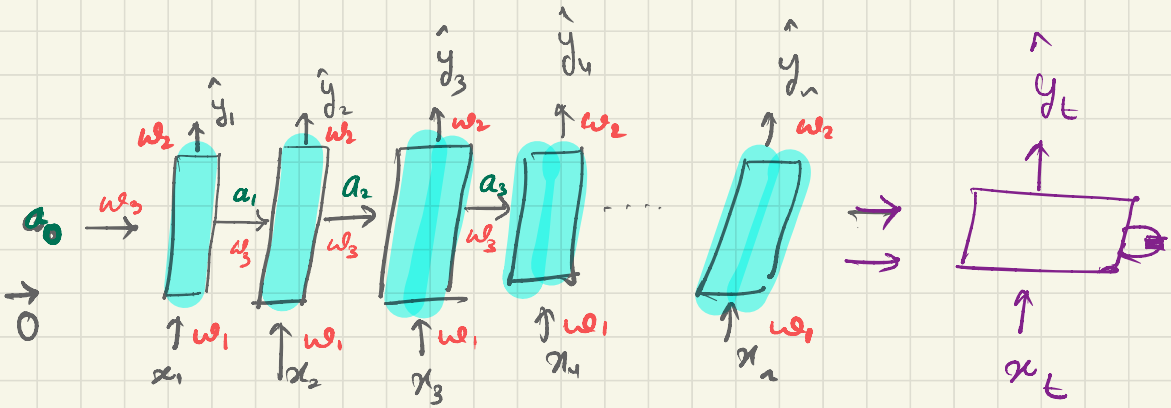
One hot encoding

model to map from X → Y.

word not in Vocabulary <unk>

→ we can perform zero padding which is in-efficient.

RNNs: Recurrent Neural Networks:



input is given from left to right

$$a_0 = \vec{0}$$

$$a_1 = g_1(w_1 \cdot x_1 + w_3 \cdot a_0 + b_1)$$

$$\hat{y}_1 = g_2(w_2 \cdot a_1 + b_2)$$

$$a_2 = g_1(w_1 \cdot x_2 + w_3 \cdot a_1 + b_1)$$

$$\hat{y}_2 = g_2(w_2 \cdot a_2 + b_2)$$

tanh / Relu : Activation.

$$a_t = g_1(w_1 \cdot x_t + w_3 \cdot a_{t-1} + b_1)$$

$$\hat{y}_t = g_2(w_2 \cdot a_t + b_2)$$

10th march 2025:

Problem with RNNs

→ Teddy Roosevelt was the president of USA.

1 1 0 0 0 0 1

→ Teddy Bears are for Sale in Big bazaar.

0 0 0 0 0 0 1

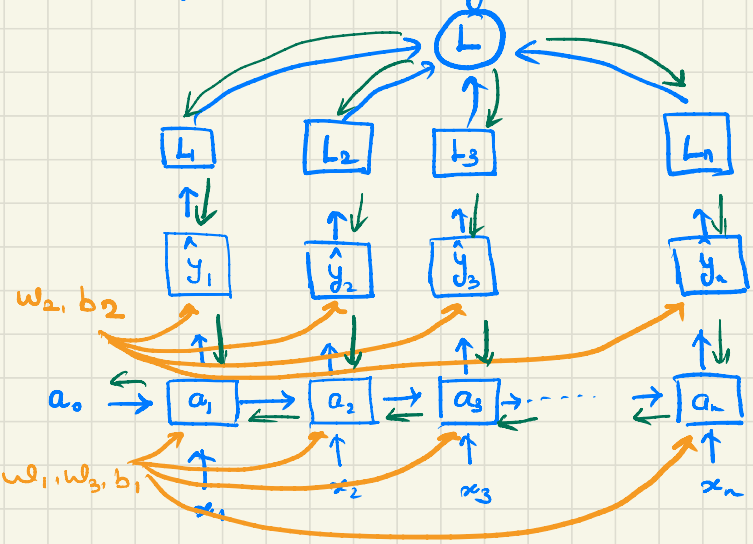
Note: to decide the NER we need to look forward for future words, which is not possible in RNNs.

Bi directional RNNs.

How to update the weights.

Computational graph

Back Propagation Through Time



Binary cross entropy loss :

$$L_1(\hat{y}_1, y_1) = -y_1 \log \hat{y}_1 - (1-y_1) \log (1-\hat{y}_1)$$

$$L(\hat{y}, y) = \sum_{i=1} L_i(\hat{y}_i, y_i)$$

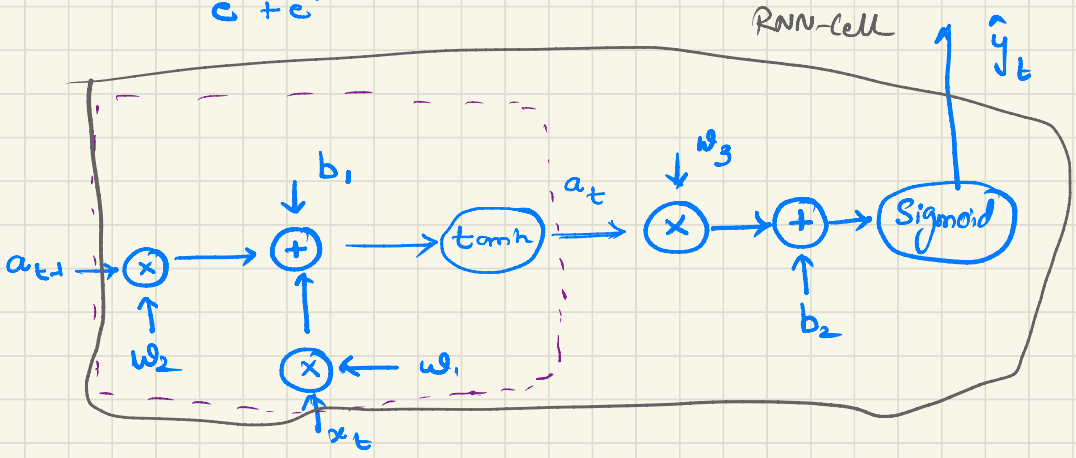
$$\begin{bmatrix} y_1 \\ 1-y_1 \end{bmatrix} \quad \begin{bmatrix} \hat{y}_1 \\ 1-\hat{y}_1 \end{bmatrix}$$

11th March 2025:

$$a_t = \tanh(w_1 x_t + w_2 a_{t-1} + b_1)$$

$$\hat{y} = \text{Sigmoid}(w_3 a_t + b_2)$$

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



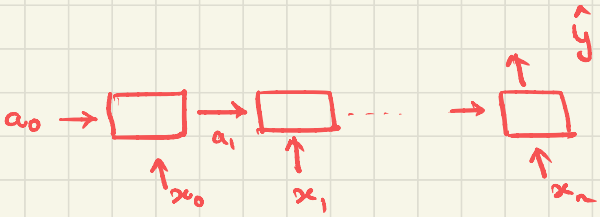
NER model we are looking is a Many to Many model. & size of input & output were same.

→ Movie Ratings

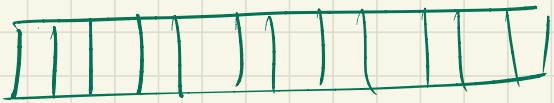
input: Review

output: number of stars $\{1, 2, 3, 4, 5\}$

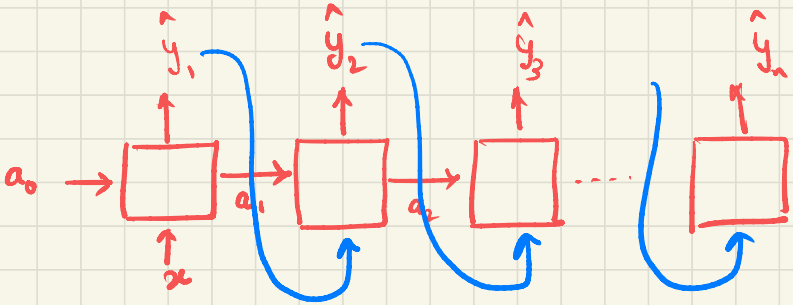
Many to One Mapping.



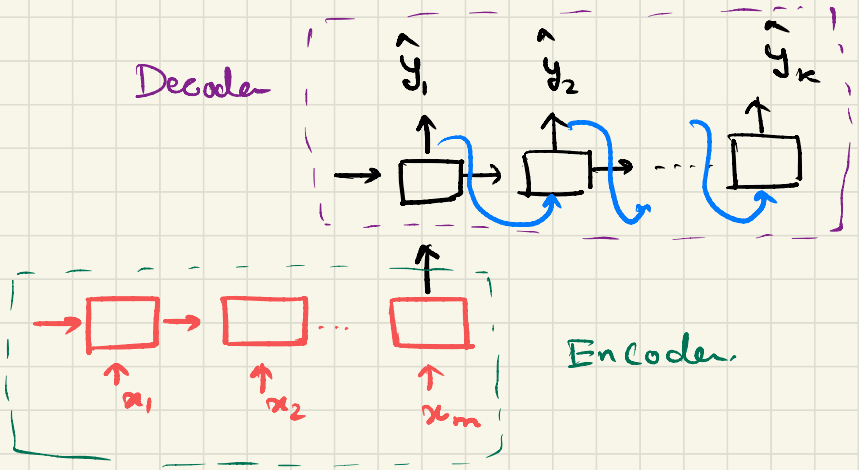
→ given the base node, play Some Raga.



One to Many Problem.



Machine translation : Many to Many. The length of i/p & o/p are diff



Consider Speech to Text

1: The apple & Pair Salad was good

2: The apple & pear Salad was good

Occurance (S_2) > Occurance (S_1)

I have total of "N" sentences in my corpus.

$$\frac{O(S_2)}{N} > \frac{O(S_1)}{N}$$

$$P[S_2] > P[S_1]$$

I like eating Pongal
x_1 x_2 x_3 x_4

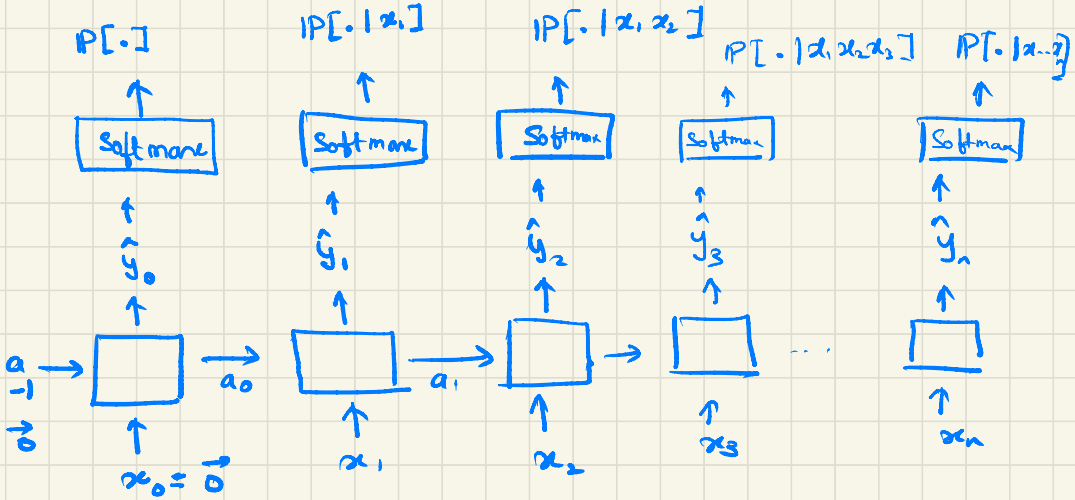
word Translation problem.

length of i/p & o/p is same.

output should choose one word out of all the words in the vocabulary.

$\therefore$ need to use Softmax.

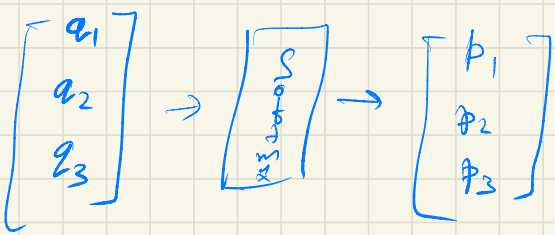
Consider there are 10 words in vocabulary then we should get a probability vector of length 10.



[Eat, the, life, is, going, on, well, orange, is, good]

The ___

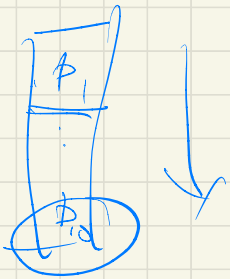
The orange ___



$$p_1 = \frac{e^{q_1}}{e^{q_1} + e^{q_2} + e^{q_3}}$$

$$p_2 = \frac{e^{q_2}}{e^{q_1} + e^{q_2} + e^{q_3}}$$

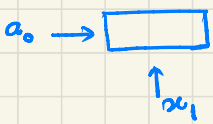
$$p_3 = \frac{e^{q_3}}{e^{q_1} + e^{q_2} + e^{q_3}}$$



12th March 2025:

Size of vocab: 5000

S_1 : I like eating dosa
S_2 : I am from mysore
S_3 : we are not in mars
S_4 : I like teaching
S_5 : Once there was a lion



→ Size of Vocabulary : 5000

each word is a vector in 5000 dim $x_1 \in \mathbb{R}^{5000}$

→ Consider the batch of size "n"

if we want to give the batch as input

then what will be the size 5000×5

input @ every time step :

→ what will be the size considering all time stamps together.

max time stamp : 10

then @ max the whole input is $5000 \times 5 \times 10$

x_1 x_2 ... x_{10}

$$a_t = (w_1 x_t + w_2 a_{t-1} + b_1)$$

$$x_t \in \mathbb{R}^{5000 \times 5}$$

$$w_1 \in ? \quad (231) \; \mathbb{R}^{10}$$

$$? \times 5000$$

$$w_1 \in \mathbb{R}$$

→ Size of w_1 & w_2 are dependent on size of a_t

→ this can be fixed by user : 231

that means for one word a_t generated is a vector of size 231

∴ for a batch 231 × 5

∴ for all the time stamps 231 × 5 × 10

$w_2 \cdot a_{t-1}$

↓ ↓

231 × 231   231 × 5

$$w_1 \in \mathbb{R}^{231 \times 5000}$$

$$w_2 \in \mathbb{R}^{231 \times 231}$$

$$a_t = w_1 * x_t + w_2 * a_{t-1} + b$$

$$= (231 * 5000)(5000 * 5) + (231 * 231)(231 * 5)$$

$$a_t = (231 * 5) + (231 * 5) + \left[\ \right]_{231*1}$$

$$\hat{y} = (w_3 * a_t + b_2)$$

$$a_t \in \mathbb{R}^{231*5}$$

$$w_3 \in \mathbb{R}^{?*231}$$

w_3 a_t

$(? \times 231)$ (231×5)

what should be the size of output

$$w_3 \in \mathbb{R}^{5000 \times 5}$$

$$b_2 \in \mathbb{R}^{5000 \times 1}$$

$$w_1 \in \mathbb{R}^{231 \times 5000} \qquad b_1 \in \mathbb{R}^{231 \times 1}$$

$$w_2 \in \mathbb{R}^{231 \times 231}$$

$$w_3 \in \mathbb{R}^{5000 \times 231} \qquad b_2 \in \mathbb{R}^{5000 \times 1}$$

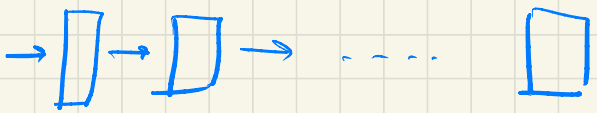
Vanishing Gradients:

next word prediction:

The man, who was . . . . . is . . . .

The men, who were . . . . . are . . . .

→ long term dependencies.



200 layer MLP.

Vanishing gradients. (NaN)