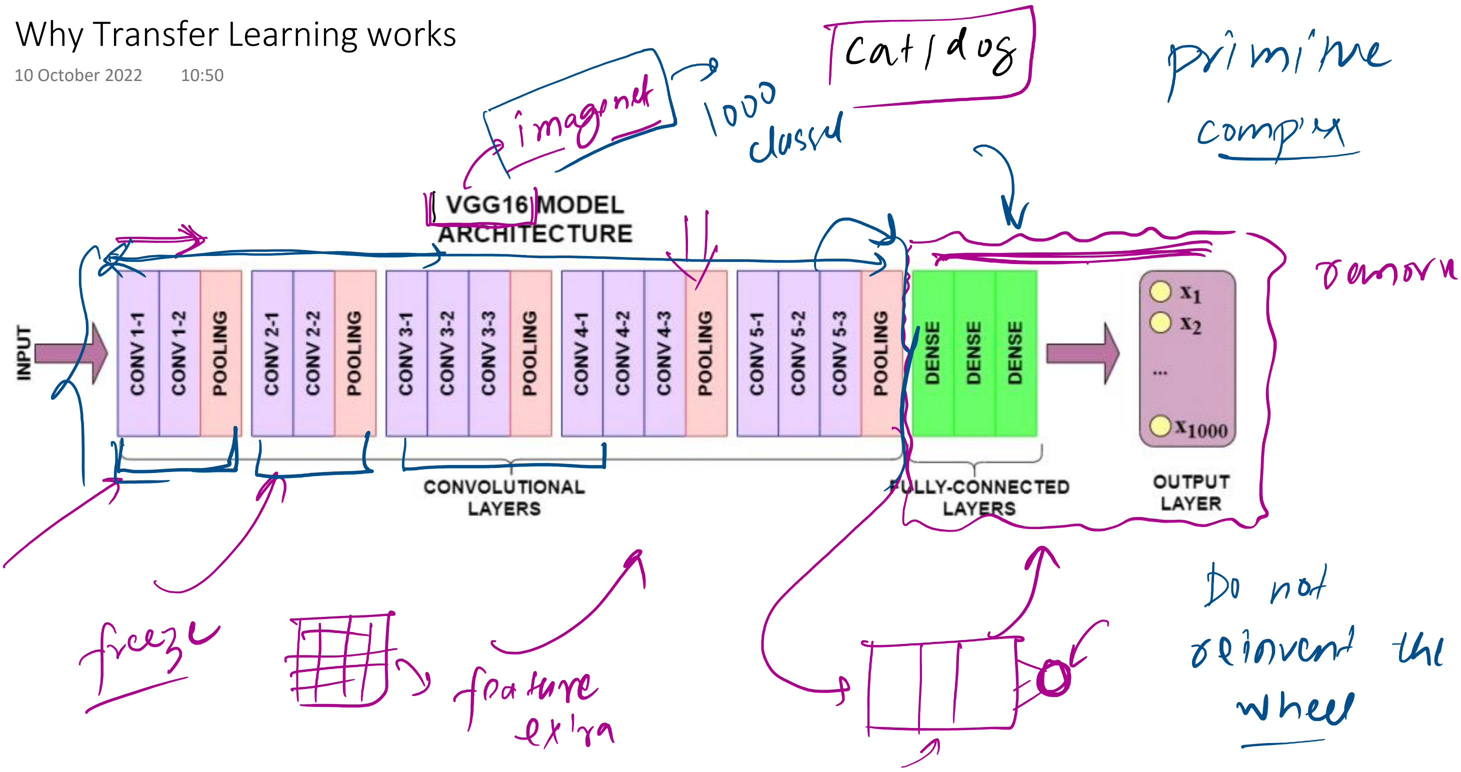
Why Transfer Learning works

10 October 2022 10:50



Ways of doing Transfer Learning

10 October 2022 10:50

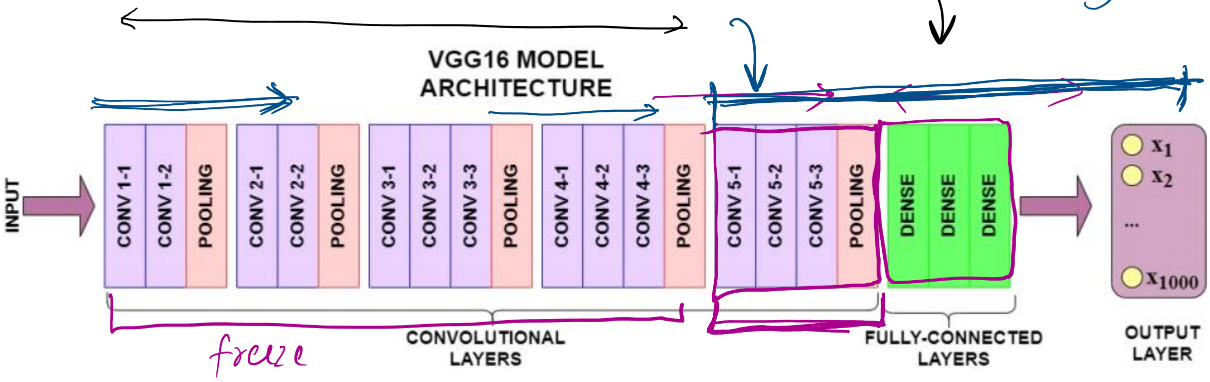
Feature extraction vs { Fine tuning } vs similar

phone / tablet

image classif → dog / cat

similar

imageNET → animals



Code

10 October 2022 10:50

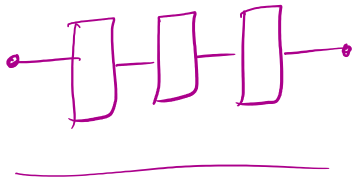
Problem with Sequential Model

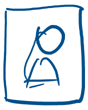
14 October 2022 16:00

ANN ↔ CNN

↳ sequential – Keras
- → 1 input
- → 1 output
- → linear

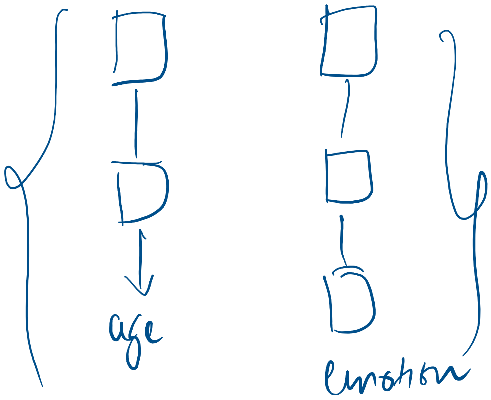
→ linear topolo

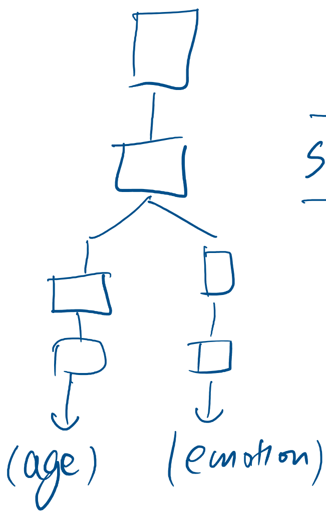


Example →  → image dataset → human faces → 25000 images

→ age?

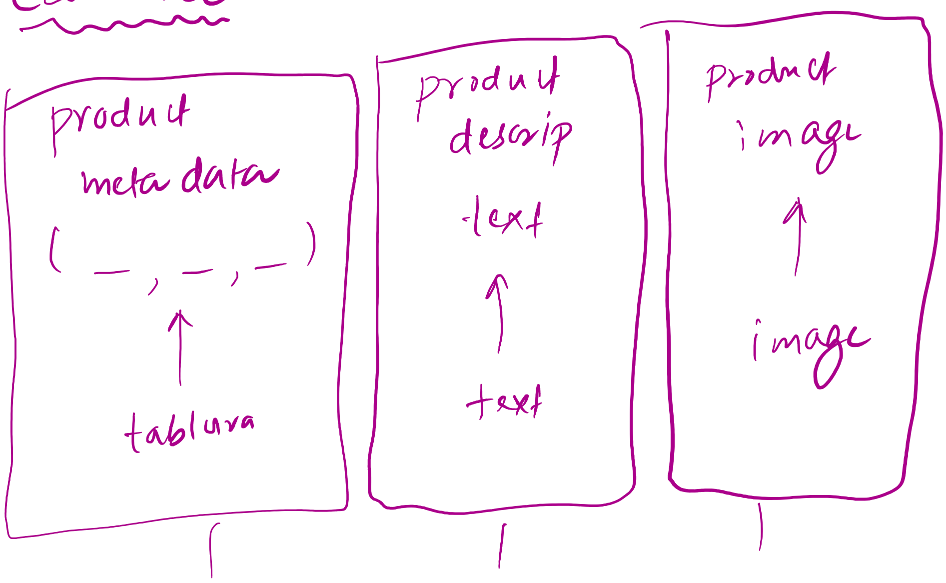
→ emotion → happy, sad, anger



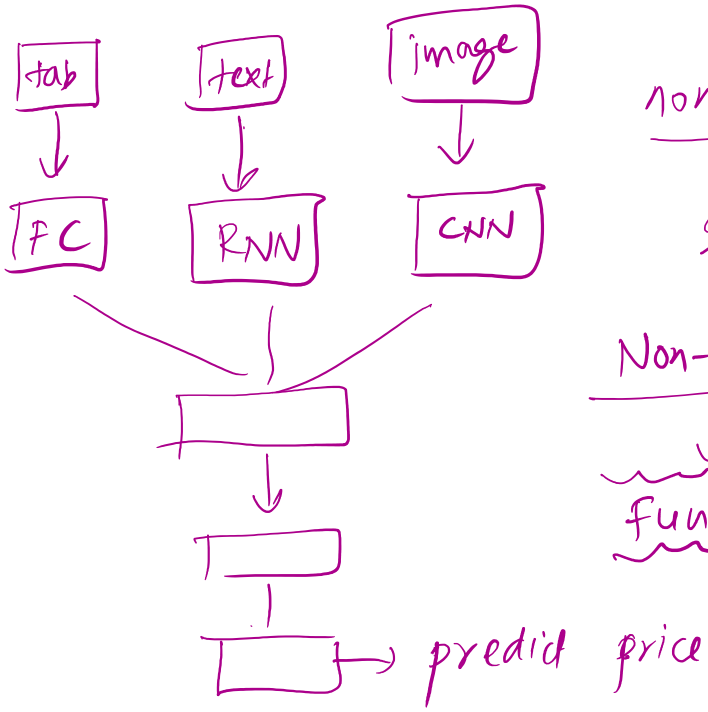


sequential API

ecommerce



predict price ?



non-linear

Sequential X

Non-linear

↓

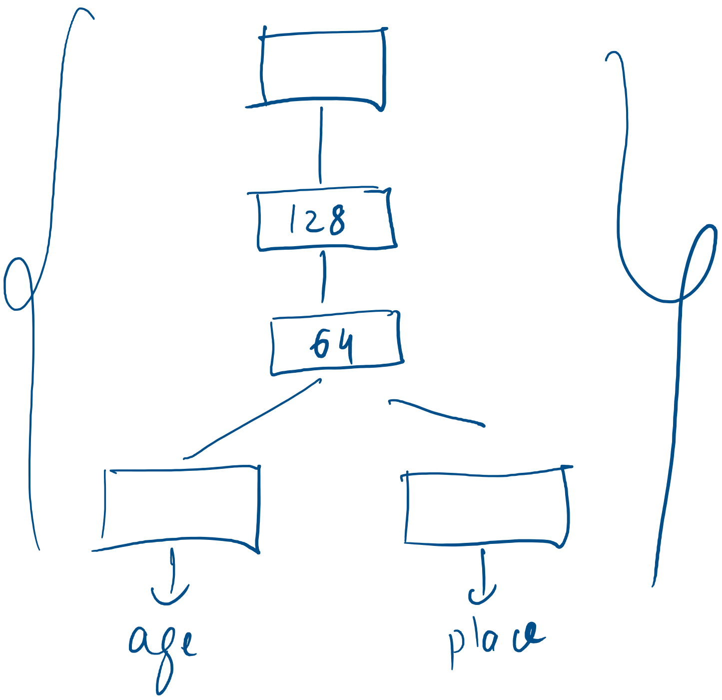
Functional API

A Simple Example

14 October 2022 16:01

| yearly salary | height | marital status |

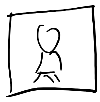
3 cols → age
→ delhi/ mumbai/



Keras functional api

Multi Output Model

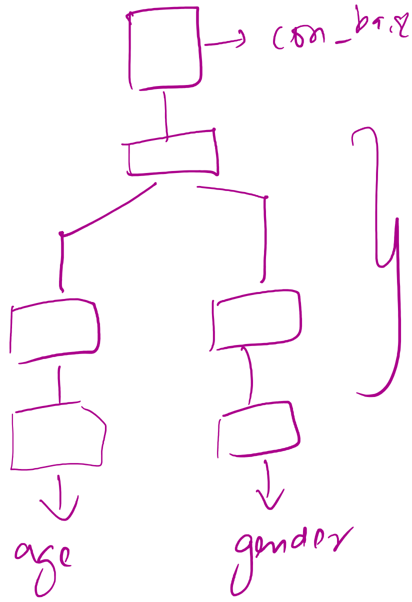
14 October 2022 16:02

age / emotion →  → 20000 → age, gender

Transfer

VGG16

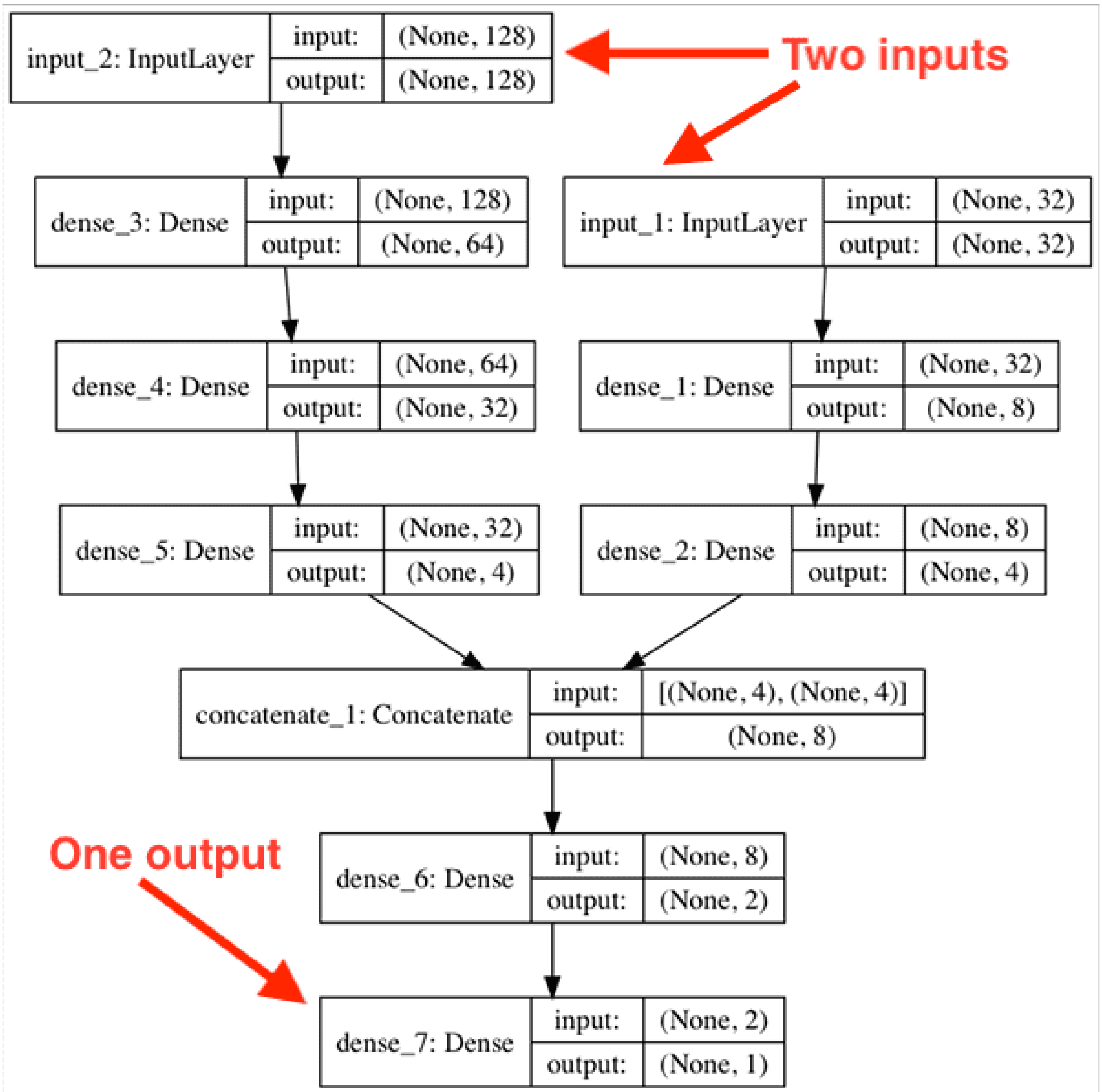
conv_base



end-end run model

Multi Input Model

14 October 2022 16:02



Shared Layers Model

14 October 2022 16:02

Sequential Data

22 October 2022 13:09

ANN → tabular data

CNN → images

RNN → Recurrent NN

is type of sequential model to work on $\boxed{\text{sequential data}}$

iq	marks	gender	placement

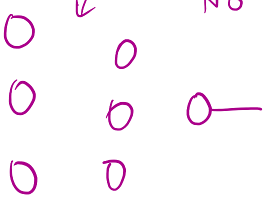
does not matter

No

iq

marks

gender.



RNN → $\boxed{\text{NLP}}$ → ML

RNN

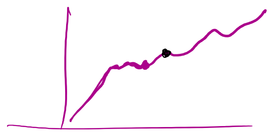
CNN → images → computer vision

eg → text → sequential data

Hi my name is Nitish

Time series

sequential



2001 2002 2003

Speech

sequential

$\boxed{\text{DNA sequence}}$

RNN

- RNN Why? → ANN / CNN
- Application → RNN
- Roadmap

Why use RNN?

22 October 2022 13:37

ANN → text classification

No

Input → 1000's

output

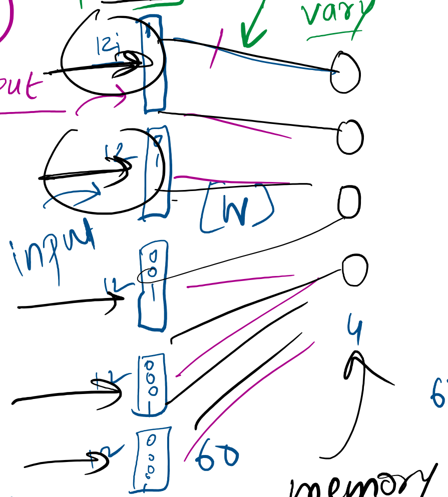
Hi my name is Nitish (5)

I love campusx (3)

India won the match (4)

↓

vectorize (OHE)



(12)

[1, 0, 0, 0, 0 . . .] [0, 1, 0 . . .]

Hi   ↑ 12 items   my

[0, 0, 1, . . .] name

12

12

12

36

48

60 × 4 = 240 weights

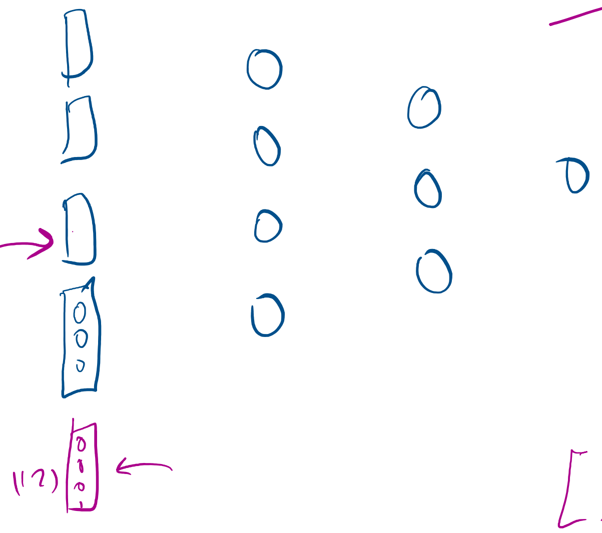
memory

textual data

diff size

zero padding

varying



Vocab → 10000 words

10k numbers

max word

100 words

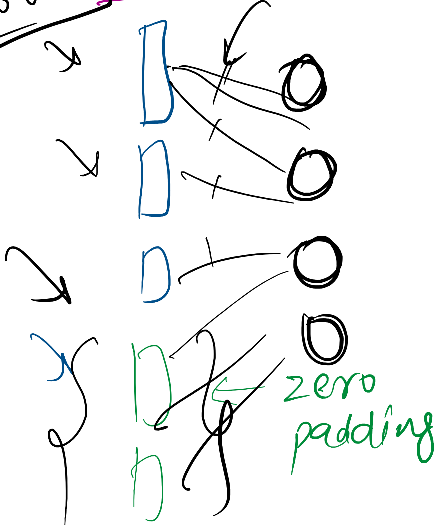
[_ _ _ _ . _]

↑

10000 numbs

100 word

min → 5 words



10000 × 100

10000000

unneccessary computation

fail

prediction

↓

200 words

no ... fore input → varying size

1) text input → varying size
2) Zero padding → unneccary computation
3) Prediction problem
4) Totally disregarding the sequence info

$\boxed{\text{RNN}} \rightarrow \boxed{\text{ANN}}$

$\uparrow$

(sequence) → $\boxed{1980}$

RNN Applications

22 October 2022 13:37

Roadmap

22 October 2022 13:37

Simple RNN → Backprop RNN → LSTM GRU → Types of RNN

→ Deep RNN → Bidirectional RNN → $\boxed{\text{seq2seq}}$

Why RNNs?

29 October 2022 13:30

zero padding → cost of compu
↑

1) sequence → any length → $\boxed{\text{ANN}}$ → fixed input size

movie reviews

$\boxed{\text{RNNs}}$
↑
NNs
↳ memory
feature

same

2) sequence contains some meaning

ANN → input

RNN architecture

RNN forward prop → prediction
↳ input → output

Codes → solidify

Data for RNN

29 October 2022 13:30

[(timesteps, input_features)]

t=1, t=2, t=3

Review	Sentiment
$\boxed{\text{movie}}$ $\boxed{\text{was}}$ $\boxed{\text{good}}$	1
movie was $\boxed{\text{bad}}$	0
movie was $\boxed{\text{not}}$ good	0

4

review 1

[[10000], [01000], [00100]]

(3,5) → # of features

no of timesteps

vocab → 5 word

movie [10000] was [01000] good [00100]

bad [00010] not [00001]

(3,5) (4,5)

Keras → SimpleRNN → batch size, timesteps, input_features

(3, 4, 5) → 3D tensor → RNN

RNN Architecture

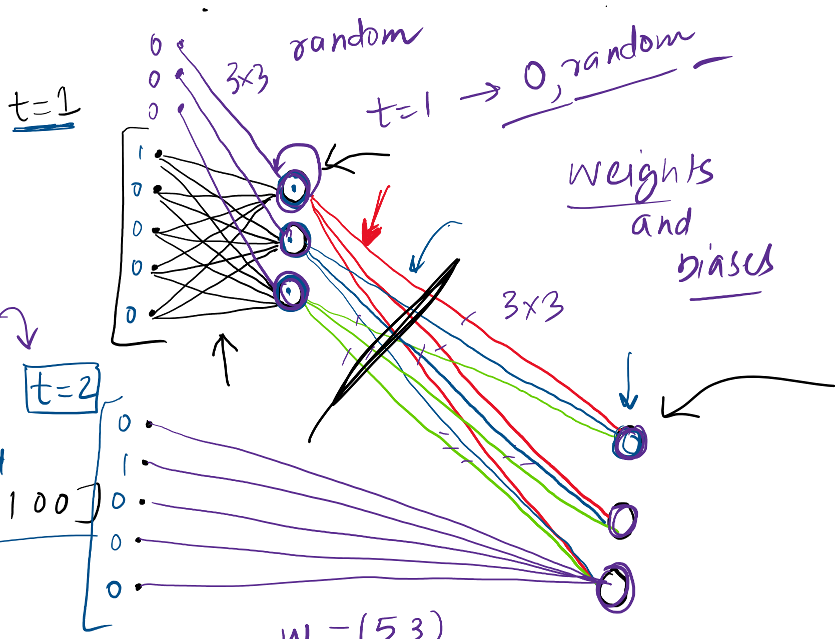
29 October 2022 13:30

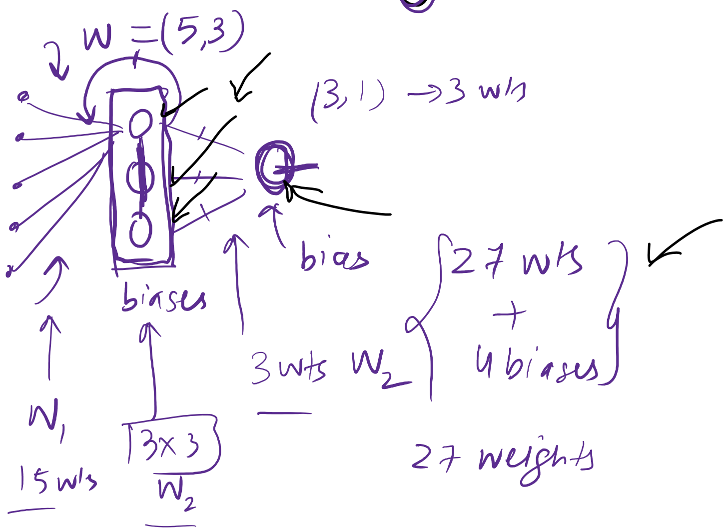
X

Review	Sentiment
X_1: movie (x_{11}) was (x_{12}) good (x_{13})	1
X_2: movie (x_{21}) was (x_{22}) bad (x_{23})	0
X_3: movie (x_{31}) was (x_{32}) not (x_{33}) good (x_{34})	0

movie [1 0 0 0 0] was [0 1 0 0 0] good [0 0 1 0 0]

bad [0 0 0 1 0] not [0 0 0 0 1]





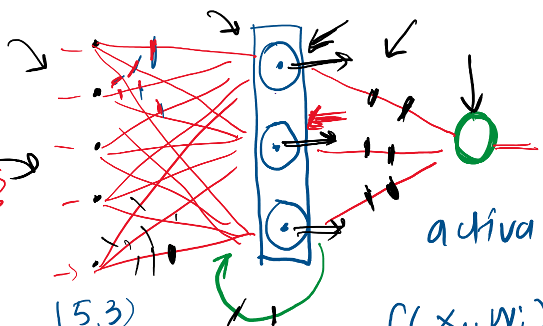
RNN Forward Prop

29 October 2022 17:02

review				sentiment
x_{11}	x_{12}	x_{13}		1
x_{21}	x_{22}	x_{23}		0
x_{31}	x_{32}	x_{33}	x_{34}	0

10 timesteps

(1,5)



(5,3)

loop

forward

unfolding through time

activation → tanh

$f(x_{11} W_i)$ → tanh / relu

→ O_1 → (1,3)

$W_h = (3,3)$

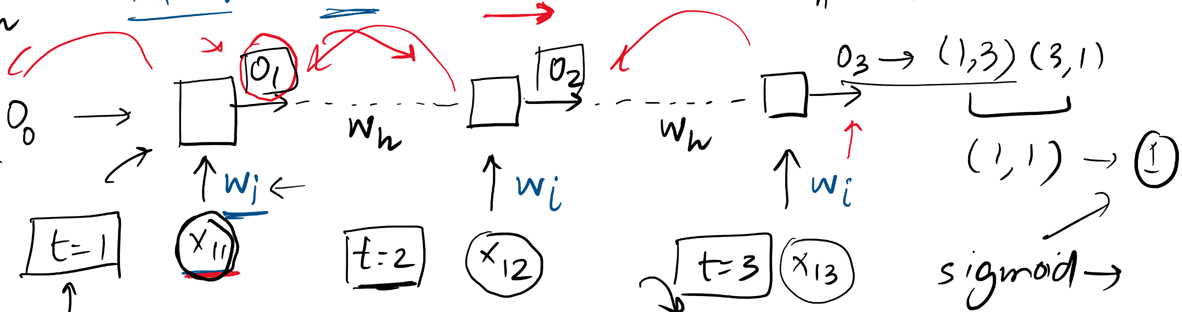
$x_{11} W_i \rightarrow (1,3)$

x_{11} → vectors 5dim

[1 0 0 0 0]

↓

one by one



(1,3) → [0 0 0]

or

random

$f(x_{11} W_i + O_0 W_h)$

x_{12}	W_i	O_1	W_h
(1,5)	(5,3)	(1,3)	(3,3)

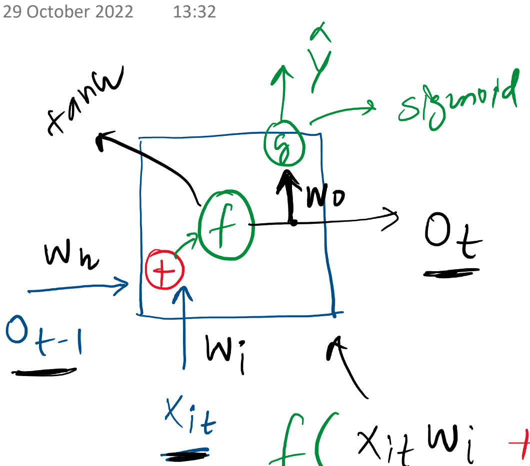
(1,3) + (1,3) → (1,3)

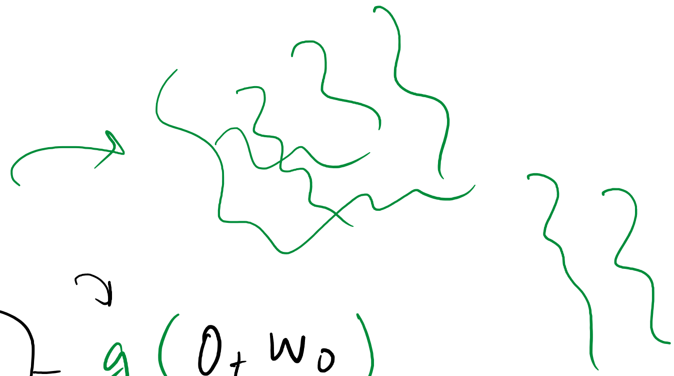
$f(\underbrace{x_{12} W_i + O_1 W_h}) =$ output

$f(x_{13} W_i + O_2 W_h)$

Simplified Representation

29 October 2022 13:32





$$\hat{y} = g(\underline{O_t W_o})$$

$$\underbrace{f(x_{it} W_i + O_{t-1} W_h)}_{O_t} \rightarrow \text{vector}$$

i = row #

t = time step

x_{11} →

Hello how are you?

Code

29 October 2022 13:32

State and Memory

29 October 2022 13:33

Keras Code Example

06 November 2022 09:14

integer encoding — embeddings

Vocab

text

Hi there
How are you

1 2 0 3 4 5

vectors

→ [1 2 0]
→ [3 4 5]

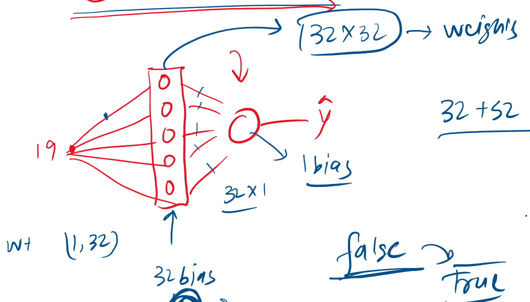
padding

RNN

Keras → text → vector
↓
RNN

50num timestep

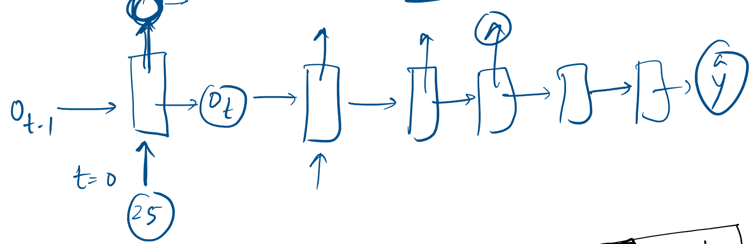
→ [25 26 18 19 . . . .]



32 + 32 + 32x32 → 1088
32
1

false → true

Named entity reco



Embedding

In natural language processing, word embedding is a term used for the representation of words for text analysis, typically in the form of a real-valued vector that encodes the meaning of the word such that the words that are closer in the vector space are expected to be similar in meaning.

this is nice

[0.7 0.1 0.3]
non-2000

dense → dimension
↓ 32 dims

semantic meaning

word → context

Sentiment

review 1 → 20 → [_ _ _ .]
review 2 → 2000

sparse repres
0
padding
sparse

20 → 1980 0 [ ]
2000 num
{1980 → 0}

→ { Word2vec
Glove }

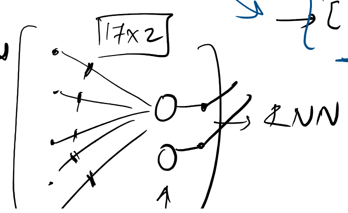
Deep learning → embedding learn → Embedding
↓
embedding ← RNN → search

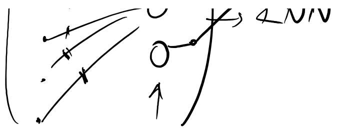
Keras embedding

[9 1 0 0 0] → 17 words
dense repres

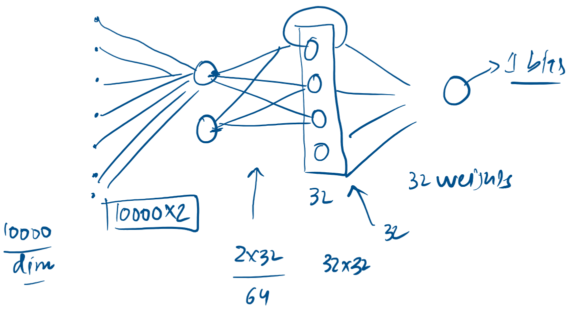
→ [x, x₂], [x, x], [x₃ x₄],
[x₅, x₆], [x₇, x₈]]

17 dim one





17 nodes



population in 9th year = x

$x + 10\%\text{ of }x = 10000$ $\quad t(n-1)$

$x + 0.1x = 10000$

$1.1x = 10000$

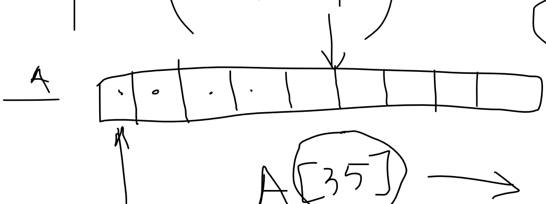
$$\frac{x-1}{x} + \frac{1}{2}\left(\frac{x-1}{x}\right)^2 + \frac{1}{2}\left(\frac{x-1}{x}\right)^3 + \frac{1}{2}\left(\frac{x-1}{x}\right)^4 + \dots$$

$\boxed{x = \frac{10000}{1.1}}$

$+ 2 +$

$+ (n-1)$

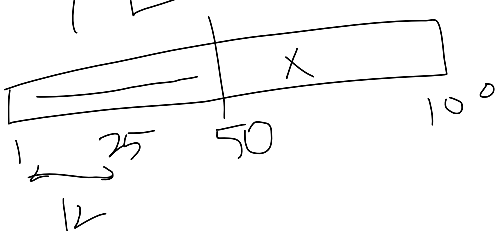
50 → 500



A[35] → t sec

A[35] →

35 $\quad \boxed{1 \times 4 \times 35}$



O(n)

O(n²) → nested loops

input → 1 10 100(x10)

time (x) 2 3 4

O(log(n))

Binary search

1 — 100

(6)

$\sqrt{(x_1}$

Binary ...

1 — 100

6

1 25 50 100

n

7

1 12 25 $\frac{n}{2}$

1 12 $\frac{n}{4}$ $\frac{n}{8}$

1 1000

O(n log n)

Sorting

$\boxed{O(2^n)}$ → exponential

1 2 3 4

2 4 8 32

2^{10}

for i in ran

O(n)

for j in (-10) n^2

O(n)

O(n+n) →

O(2n)

→ O(n)

$O(\cancel{n}+n^2)$ $O(n^2)$

for i in arrs

for j in

0 1 2 3 4 8 6 7 5 n^2)

25 → '25'

str( )

n = 345 % 10

digits[5]

5 + ''

= '5'

345//10 → 34 != 0

34 % 10 → 4

'' + '5' .

$34 \% 10 \rightarrow 4$

$4 + '5' \rightarrow$ 45

$34 // 10 \rightarrow 3$

$3 != 0$

$3 \% 10 \Rightarrow 3$

digits[3]

$3 + '45'$
$= '345'$

Loop $n \rightarrow 10$ times

$\boxed{345} = n \Rightarrow 3450$

$3 \rightarrow 4$

inp $\rightarrow$ 10
time $\rightarrow$ 1 $\rightarrow$ log

$O(n)$
$O(n)$ $\rightarrow O(n+n)$
$O(2n) \rightarrow O(n)$

$O(n)$
$O(n)$ $\rightarrow O(n*n)$
$O(n^2)$

$O(n)$
$O(n)$
$O(1000000)$

n^2 $\boxed{1000000}$
×
$O(n^2)$

$1 \rightarrow \frac{n}{2}$ len(L)/2

$4 \frac{n}{2}$ 2n

$O(n)$ $n=100$

$\boxed{50, 100}$ $\boxed{\frac{n}{2}}$

$2 \rightarrow \boxed{100 = n}$ $\boxed{\frac{n}{2}}$

$j=1 \rightarrow 2$
$j=2 \rightarrow 4$ $\boxed{2-100}$
$j=3 \rightarrow 8$
$j=4 \rightarrow 32$

$\frac{n}{\cancel{2}} \times \log n$
$\boxed{n \log n}$

$\boxed{O(1)}$ → constant

n = 345

3+4+5 → 12

5

3 4   4

3   3

$\boxed{0}$

3450 → 4

$\boxed{\log}$

inp → 10   100

out → 1   2   3

fibonacci

↳ function

↳ recursion

0 1 1 2 3 5 8 13

5

n=1   n=0 → 1

fib(n)

function calls

input → n → 10

#fcalls →

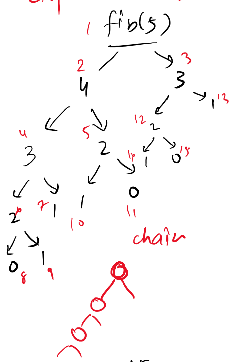
fib(3) 1   n=3

fib(2) 2   fib(1) 3 → 1

fib(1)   fib(0)

1 4   1 5

exponential → bad



$O(2^n)$ 15 20   50

input 1 2 3 4

.. 2 4 8 16

$O(2^n)$ 50

input 1 2 3 4

ti 2 4 8 16

$\boxed{n = 50, 100, 500}$

↑ weeks

exponential

↳ days/ weeks

subset → O(?)

power set

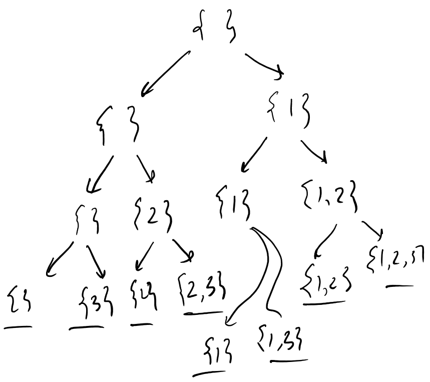
$\{1,2\} \rightarrow \{\{\}, \{1\}, \{2\}, \{1,2\}\}$

$\{1,2,3\}$ ↓

$\{\}, \{1\}, \{2\}, \{3\}, \{1,2\} \{2,3\}$

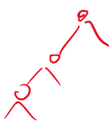
$\{3,1\}$ $\{1,2,3\}$

$\{1,2,3\}$



reduce → divide → log

Increa → mult → exp



→ exponential

$2^2 = 4$

$\{1,2\} \rightarrow 4$

$2^3 = 8$

$\{1,2,3\} \rightarrow 8$

$2^4 = 16$

$O(2^n)$

O(?)

$$T(n) = \begin{cases} 3T(n-1) & \text{if } n>0 \\ 1, & \text{otherwise} \end{cases}$$

$n > 0$

$$T(n) = 3T(n-1)$$
$$= 3[3T(n-2)]$$
$$= 3^2 T(n-2)$$
$$= 3^2[3T(n-3)]$$

$= 3^3 T(n-3)$

$= 3^n T(n-n)$

$= 3^n \underline{T(0)}$

$T(n) = \boxed{3^n} \rightarrow O(3^n)$

$$T(n) = \begin{cases} 2T(n-1)-1 & \text{if } n>0 \\ \underline{1}, & \text{otherwise} \rightarrow \underline{\text{constant}} \end{cases}$$

$T(n) = 2\underline{T(n-1)} - 1$

$= 2[2T(n-2)-1]-1$

$= 2^2 \underline{T(n-2)} - 2 - 1$

$= 2^2[2T(n-3)-1] - 2 - 1$

$= 2^3 T(n-3) - 2^2 - 2^1 - 2^0$

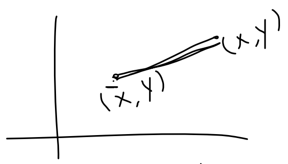
$= \underline{2^n} \underline{T(n-n)} - 2^{n-1} - 2^{n-2} \ldots - 2^1 - 2^0$

$= 2^n - [2^{n-1} + 2^{n-2} + \ldots + 2^1 + 2^0]$

$= 2^n - [2^n - 1] = \cancel{2^n} - \cancel{2^n} + 1$

$O(1) \rightarrow$ constanst

$\sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$



$\boxed{H = 4}$ $\boxed{L = 10}$

$\boxed{D = 4 - C}$

$D = ?$ $C = ?$

$\boxed{C + D = 4}$ 1

$2C + 4D = 10$

15 5 $4^2 + 5^2$

2 6 9 12 15

$1^2 + 2^2 + 3^2 + 4^2 + 5^2$

$n = 5$

$[3, 6 \to 5^{+n}]$

$a = 3 \quad n = 5$

$d = 6 - 3$

1009

2009

$\frac{2}{3} + \frac{4}{5} = \boxed{\frac{10+12}{15}} \to \frac{22}{15}$

0 1 1 2 3 5

$5! = 5 \times 4 \times 3 \times 2 \times 1$

$\boxed{1000}$

#

1002

$\boxed{2222}$

UP → 1
D − 3
R − 4
L − 3

a=[1,2]

a=b

a=b[:]

a →

b →

Till Now

17 November 2022 17:01

→ Roadmap → overview

↳ RNN → theoritical

↳ RNN → keras → sentiment analy

↳ Types of RNN → Backprop RNN

Many to One

One to many

Many 2 many

One 2 one

examples

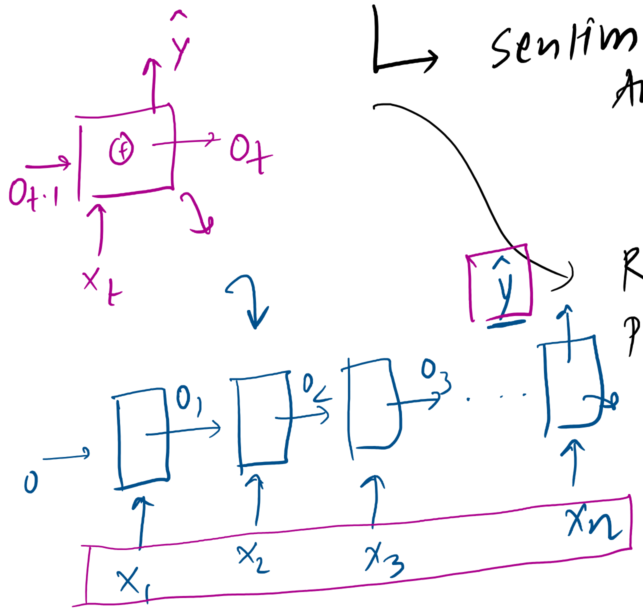
Many to One

17 November 2022 17:02

input → sequence → sent, chars, time series

out → non-sequential → int/num → scalars → (1,0)

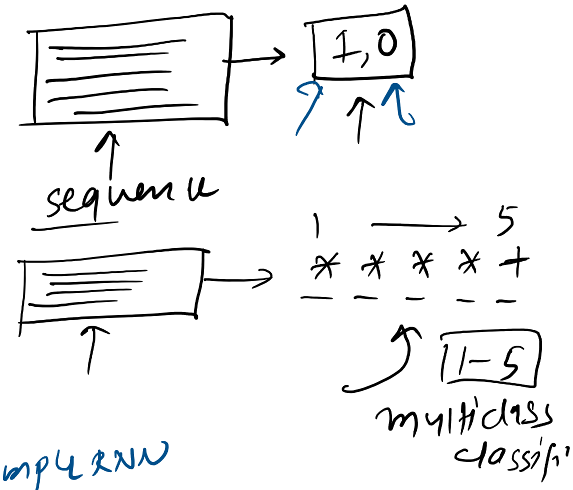
↳ sentiment → Analysis



Rating prediction

recurrent → simpleRNN

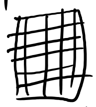
many to one



One to Many

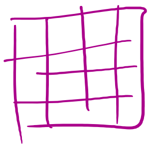
17 November 2022 17:02

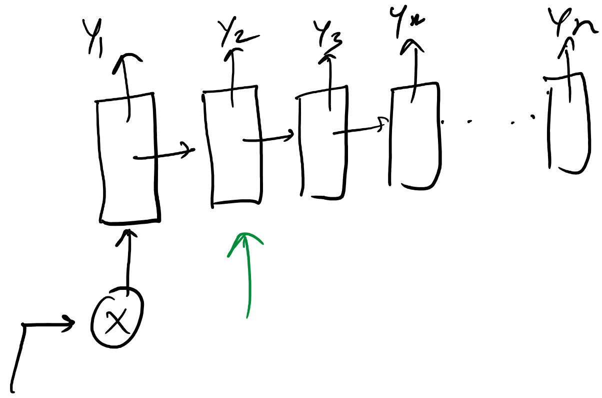
→ normal non sequential

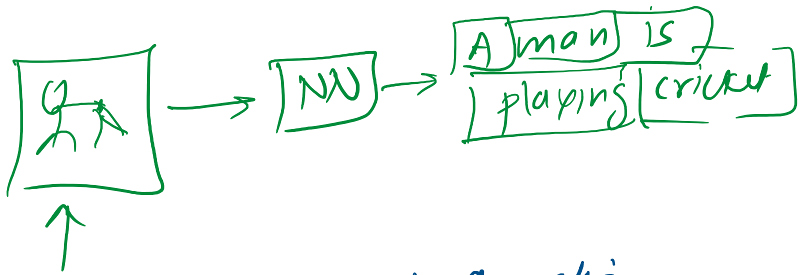
↳ 

→ Output → sequentials

image captioning

 → textual repres





→ music generation

Many to Many

17 November 2022 17:02

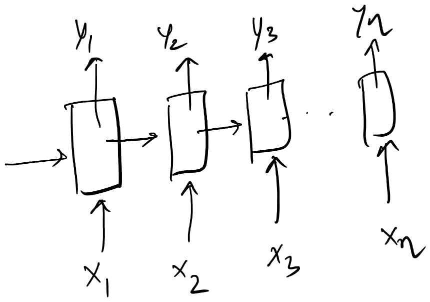
input → sequent
out → sequen } → seq2seq

Same length | Variable length

input seq == output seq

pos tagging → NLP

1	2	3	4
My	name	is	Nitish
↓	↓	↓	↳
pronoun	noun	verb	Noun
1	2	3	4



same length many2many RNN

NER

Lets meet at 7pm at the airport

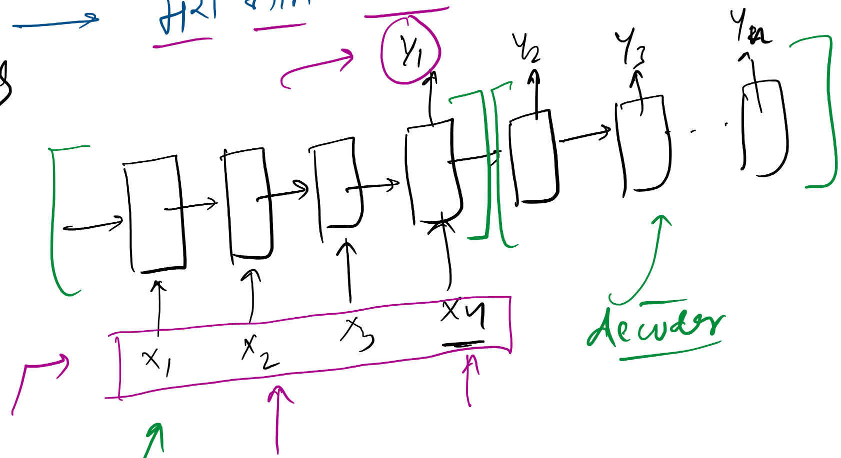
Variable length

machine trans
↳ 1 lang → 2 lang
↓
google translate

encoder decoder

→ Hi my name is Nitish

→ मेरा नाम नीतीश है



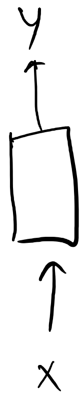
encoder

One to One

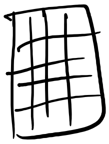
17 November 2022 17:02

one → non sequential → image classifi

out → non seque



→ NN → ANN
↘ CNN



→ 1,0

- many 2 one
- one 2 many
- many 2 many
  - same
  - variable

Backpropagation in RNN

01 December 2022 16:43

RNN intro → RNN → practical ↓

types of RNN

→ RNN → Backprop → $\boxed{\text{BPTT}}$

Many to One RNN

Sentiment Analysis

text → 1/0

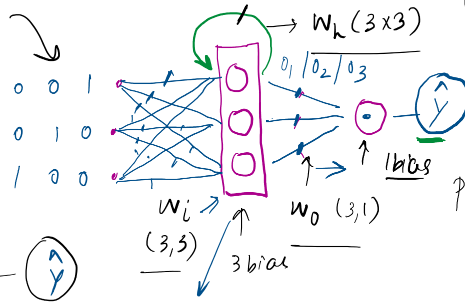
text	
cat mat rat	1
rat rat mat	1
mat mat cat	0

Vocab → Cat Mat Rat
[1 0 0] [0 1 0] [0 0 1]

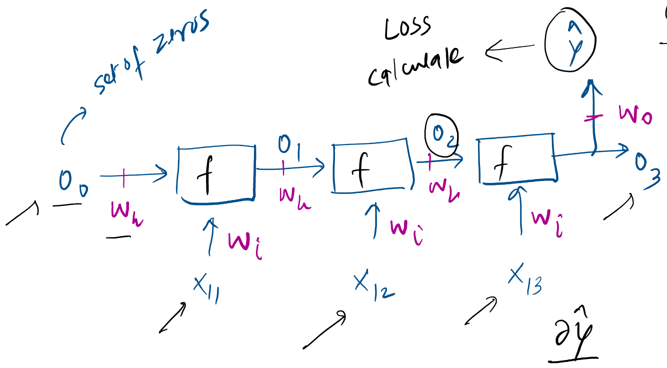
	o_1	o_2	o_3	$\hat{y}$	y
X_1	[1 0 0]	[0 1 0]	[0 0 1]		1
X_2	[0 0 1]	[0 0 1]	[0 1 0]		1
X_3	[0 1 0]	[0 1 0]	[1 0 0]		0

$\hat{y} \to L$

$\boxed{W_i \quad W_h \quad W_o}$



forward prop



$$o_1 = f(X_{11} W_i + o_0 W_h)$$

$$o_2 = f(X_{12} W_i + o_1 W_h)$$

$$o_3 = f(X_{13} W_i + o_2 W_h)$$

$\dfrac{\partial \hat{y}}{\partial W_o}$

y → $\hat{y} = \sigma(o_3 W_o)$

loss calculate → minimize

$$L = -y_i \log \hat{y}_i - (1-y_i)\log(1-\hat{y}_i)$$

L min   gradient descent

$\dfrac{\partial L}{\partial \hat{y}}$

$W_i \quad W_h \quad W_o$

$$W_i = W_i - \eta \boxed{\frac{\partial L}{\partial W_i}}$$

$$W_h = W_h - \eta \boxed{\frac{\partial L}{\partial W_h}}$$

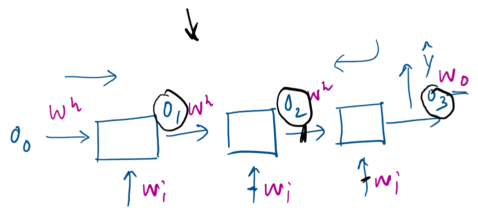
$$W_o = W_o - \eta \boxed{\frac{\partial L}{\partial W_o}}$$

Backprop

unfold → 3 times

$\dfrac{\partial L}{\partial W_o}$

$L \to \hat{y} \to o_3,\ W_o$

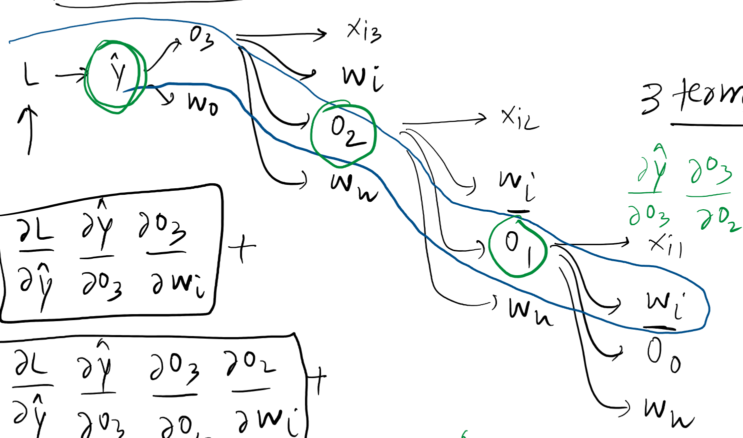


$\frac{\partial L}{\partial w_o}$ $L \to \hat{y} \to o_3, w_o$

$$\boxed{\frac{\partial L}{\partial w_o} = \frac{\partial L}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial w_o}}$$

$\boxed{\frac{\partial L}{\partial w_o}}, \boxed{\frac{\partial L}{\partial w_i}}, \boxed{\frac{\partial L}{\partial w_h}}$

$\frac{\partial L}{\partial w_i}$



3 terms

$\frac{\partial \hat{y}}{\partial o_3} \frac{\partial o_3}{\partial o_2}$

$o_1 = f(x_{i1} w_i + o_0 w_h)$

$o_2 = f(x_{i2} w_i + o_1 w_h)$

$o_3 = f(x_{i3} w_i + o_2 w_h)$

$\hat{y} = \sigma(o_3 w_o)$

$$\boxed{\frac{\partial L}{\partial w_i}} = \boxed{\frac{\partial L}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial o_3} \frac{\partial o_3}{\partial w_i}} + \boxed{\frac{\partial L}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial o_3} \frac{\partial o_3}{\partial o_2} \frac{\partial o_2}{\partial w_i}} + \boxed{\frac{\partial L}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial o_3} \frac{\partial o_3}{\partial o_2} \frac{\partial o_2}{\partial o_1} \frac{\partial o_1}{\partial w_i}}$$

10 words

10 terms

$t=1$

$$\frac{\partial \hat{y}}{\partial o_1} = \frac{\partial \hat{y}}{\partial o_3} \frac{\partial o_3}{\partial o_2} \frac{\partial o_2}{\partial o_1}$$

$$\boxed{\frac{\partial L}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial o_3} \frac{\partial o_3}{\partial w_i}}$$

$$\frac{\partial L}{\partial w_i} = \sum_{j=1}^{3} \frac{\partial L}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial o_j} \frac{\partial o_j}{\partial w_i}$$

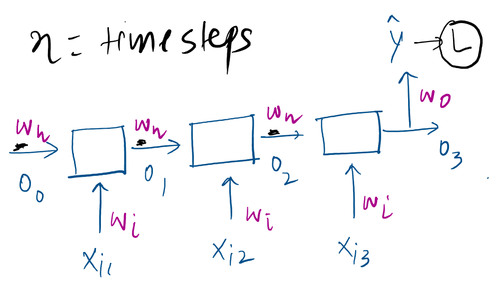
$$= \frac{\partial L}{\partial \hat{y}} \boxed{\frac{\partial \hat{y}}{\partial o_1}} \frac{\partial o_1}{\partial w_i} = \boxed{\frac{\partial L}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial o_3} \frac{\partial o_3}{\partial o_2} \frac{\partial o_2}{\partial o_1} \frac{\partial o_1}{\partial w_i}}$$

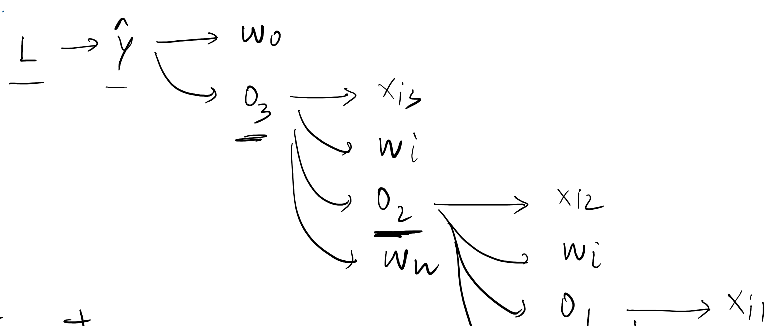
$$\frac{\partial L}{\partial \hat{y}} \boxed{\frac{\partial \hat{y}}{\partial o_2}} \frac{\partial o_2}{\partial w_i} = \boxed{\frac{\partial L}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial o_3} \frac{\partial o_3}{\partial o_2} \frac{\partial o_2}{\partial w_i}}$$

$$\boxed{\frac{\partial L}{\partial w_i} = \sum_{j=1}^{n} \frac{\partial L}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial o_j} \frac{\partial o_j}{\partial w_i}}$$

n = timesteps

$$\boxed{\frac{\partial L}{\partial w_h}}$$





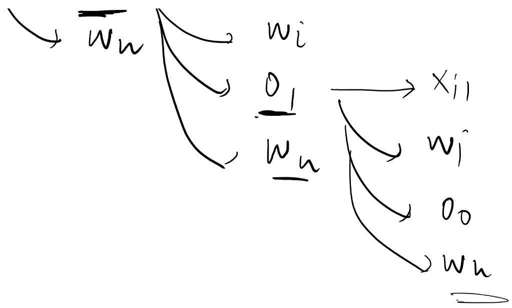
$$\frac{\partial L}{\partial w_h} = \frac{\partial L}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial o_3} \frac{\partial o_3}{\partial w_h} +$$

$$\partial L \quad \partial \hat{y} \quad \partial o_2 \quad \partial o_2 \quad +$$

$$\partial W_h \quad \partial y \quad \partial o_3 \quad \partial W_h$$

$$\frac{\partial L}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial o_3} \frac{\partial o_3}{\partial o_2} \frac{\partial o_2}{\partial W_h} +$$

$$\frac{\partial L}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial o_3} \frac{\partial o_3}{\partial o_2} \frac{\partial o_2}{\partial o_1} \frac{\partial o_1}{\partial W_h}$$



$$\boxed{\frac{\partial L}{\partial W_h} = \sum_{j=1}^{n} \frac{\partial L}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial o_j} \frac{\partial o_j}{\partial W_h}}$$

n = timesteps

for $j=3$

$$\frac{\partial L}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial o_3} \frac{\partial o_3}{\partial W_h} \rightarrow \frac{\partial L}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial o_3} \frac{\partial o_3}{\partial o_1} \frac{\partial o_1}{\partial W_h}$$

for $j=10$

$$\frac{\partial L}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial o_{10}} \frac{\partial o_{10}}{\partial o_1} \frac{\partial o_1}{\partial W_h} \qquad \boxed{\prod_{t=2}^{j} \frac{\partial o_t}{\partial o_{t-1}}}$$

$$\prod_{t=2}^{t=3} \frac{\partial o_t}{\partial o_{t-1}} = \frac{\partial o_2}{\partial o_1} \frac{\partial o_3}{\partial o_2}$$

$$o_t = f(x_{it} W_{inp} + o_{t-1} W_h)$$

$$\frac{\partial o_t}{\partial o_{t-1}} = \prod_{t=2}^{j} \underbrace{f'(x_{it} W_{inp} + o_{t-1} W_h)}_{[0-1]} W_h$$

Problem with RNN

19 December 2022 16:33

RNNs → sequential data → textual, time series

suffer 2 major problems → LTSMs Why

RNN

1 → Problem of long term denpendency → Unstable gradients

→ ~~Unstable gradients~~ Stagnated values

Application next word predicts

timesteps

→ RNN

→ Marathi is spoken in Mahar

short

RNN

→ Maha is a beautiful place. I went there last year, but I could not enjoyed properly b/c a I dont understand marathi

Vanishing gradient prob → long term

→ RNN ↘ keypad

Exploding gradient problem → RNN

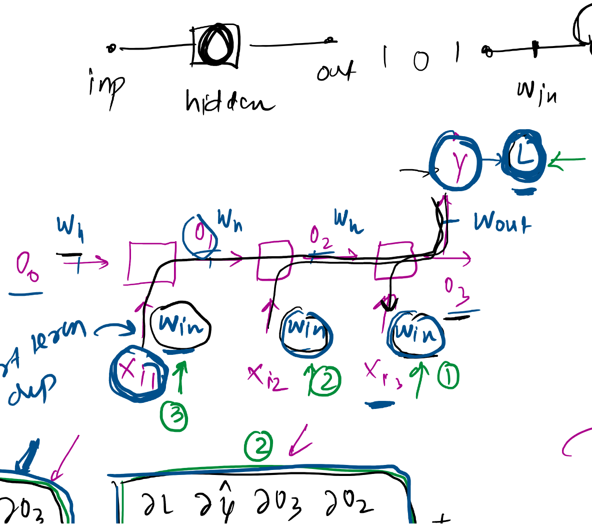
possible sol^n ↳ remove

Problem # 1 → Problem of long term dependency → Vanishing

Inp			out
1	0	1	1
0	0	1	0
0	0	0	0
1	1	1	1

3 time stamps

0 - 1



L → min

W_h, W_{in}, W_{out}

↓

gradient des

$$W_{in} = W_{in} - \eta \frac{\partial L}{\partial W_{in}}$$

$$W_{out} = W_{out} - \eta \frac{\partial L}{\partial W_{out}}$$

$\partial L \quad \partial \hat{y} \quad \partial O_3$ $\quad$ $\partial L \quad \partial \hat{y} \quad \partial O_3 \quad \partial O_2$

$$\frac{\partial L}{\partial w_{in}} = \frac{\partial L}{\partial \hat{y}}\frac{\partial \hat{y}}{\partial o_3}\frac{\partial o_3}{\partial w_{in}} + \frac{\partial L}{\partial \hat{y}}\frac{\partial \hat{y}}{\partial o_3}\frac{\partial o_3}{\partial o_2}\frac{\partial o_2}{\partial w_{in}} + \frac{\partial L}{\partial \hat{y}}\frac{\partial \hat{y}}{\partial o_3}\frac{\partial o_3}{\partial o_2}\frac{\partial o_2}{\partial o_1}\frac{\partial o_1}{\partial w_{in}}$$

1 2 3

long term dep

inf

$$w_h = w_h - \eta \frac{\partial L}{\partial w_h}$$

→ 3

100 time steps

$$\frac{\partial L}{\partial \hat{y}}\frac{\partial \hat{y}}{\partial o_{100}}\frac{\partial o_{100}}{\partial o_{99}} \cdots \cdots \frac{\partial o_2}{\partial o_1}\frac{\partial o_1}{\partial w_{in}} \rightarrow \frac{\partial L}{\partial \hat{y}}\frac{\partial \hat{y}}{\partial o_{100}}\prod_{t=2}^{100}\left(\frac{\partial o_t}{\partial o_{t-1}}\right)\frac{\partial o_1}{\partial w_{in}}$$

$$\frac{\partial o_2}{\partial o_1}\frac{\partial o_3}{\partial o_2}\frac{\partial o_4}{\partial o_3}\cdots\frac{\partial o_{100}}{\partial o_{99}}$$

$$o_1 = \tanh(x_{i1} w_{in} + o_0 w_h)$$

$$o_t = \tanh(x_{it} w_{in} + o_{t-1} w_h)$$

100 or more

$$\frac{\partial L}{\partial \hat{y}}\frac{\partial \hat{y}}{\partial o_{100}}\prod_{t=2}^{100}\left(\tanh'(\_)\, w_h\right)\frac{\partial o_1}{\partial w_{in}} \simeq 0 \simeq 0$$

0-1 0-1

$$\frac{\partial o_t}{\partial o_{t-1}} = \tanh'(x_{it} w_{in} + o_{t-1} w_h)\, w_h$$

relu

0-1 0-1 (1)

Vanishing grad.

Identity matrix

Sol^n

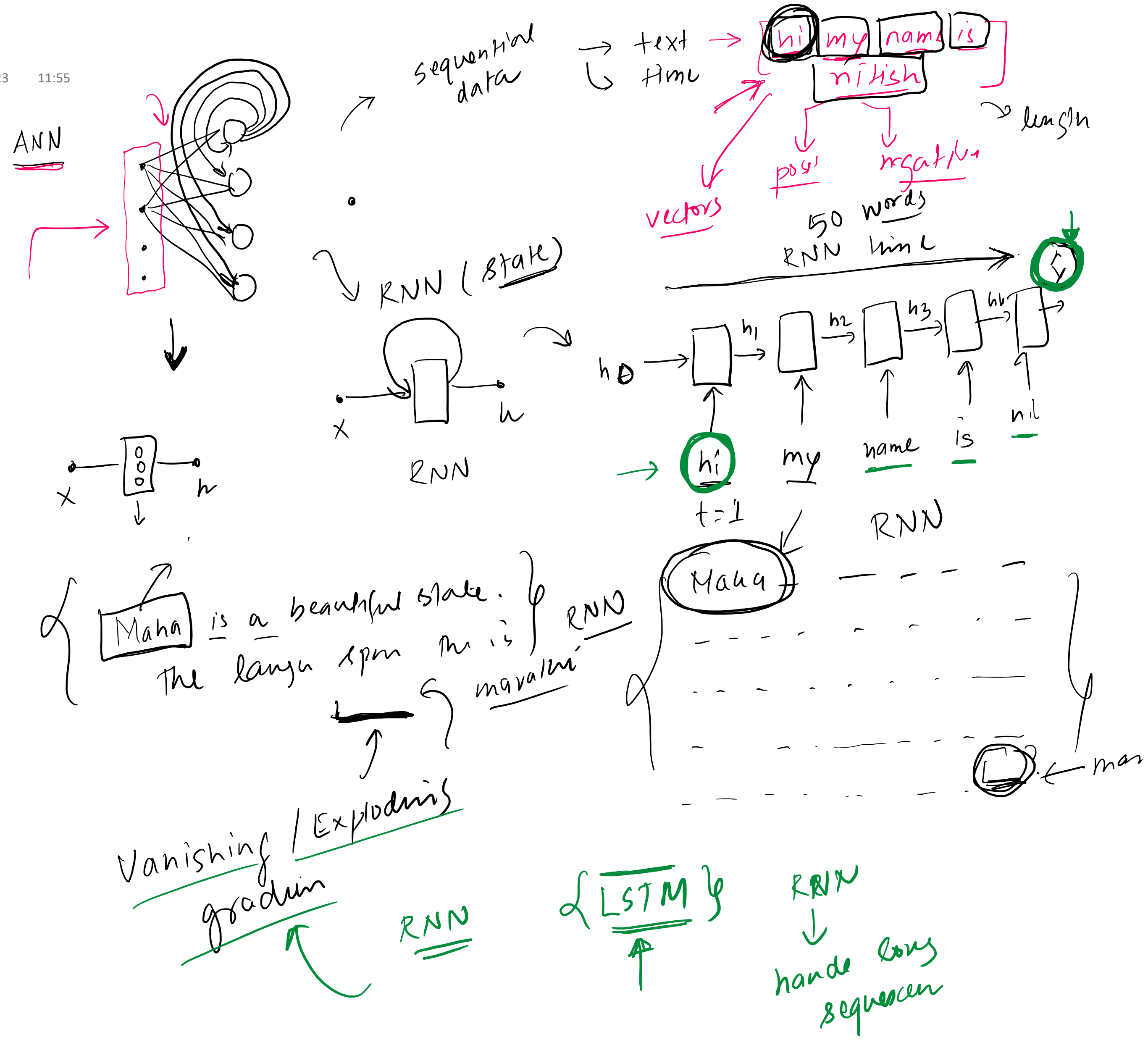
1) Diff activation → relu / leaky relu
2) Better weight init
3) Skip conns
4) LSTM

Problem #2 → Unstable Training (Exploding gradients)

1) Gradient Clipping
2) Controlled learning rate
3) LSTM

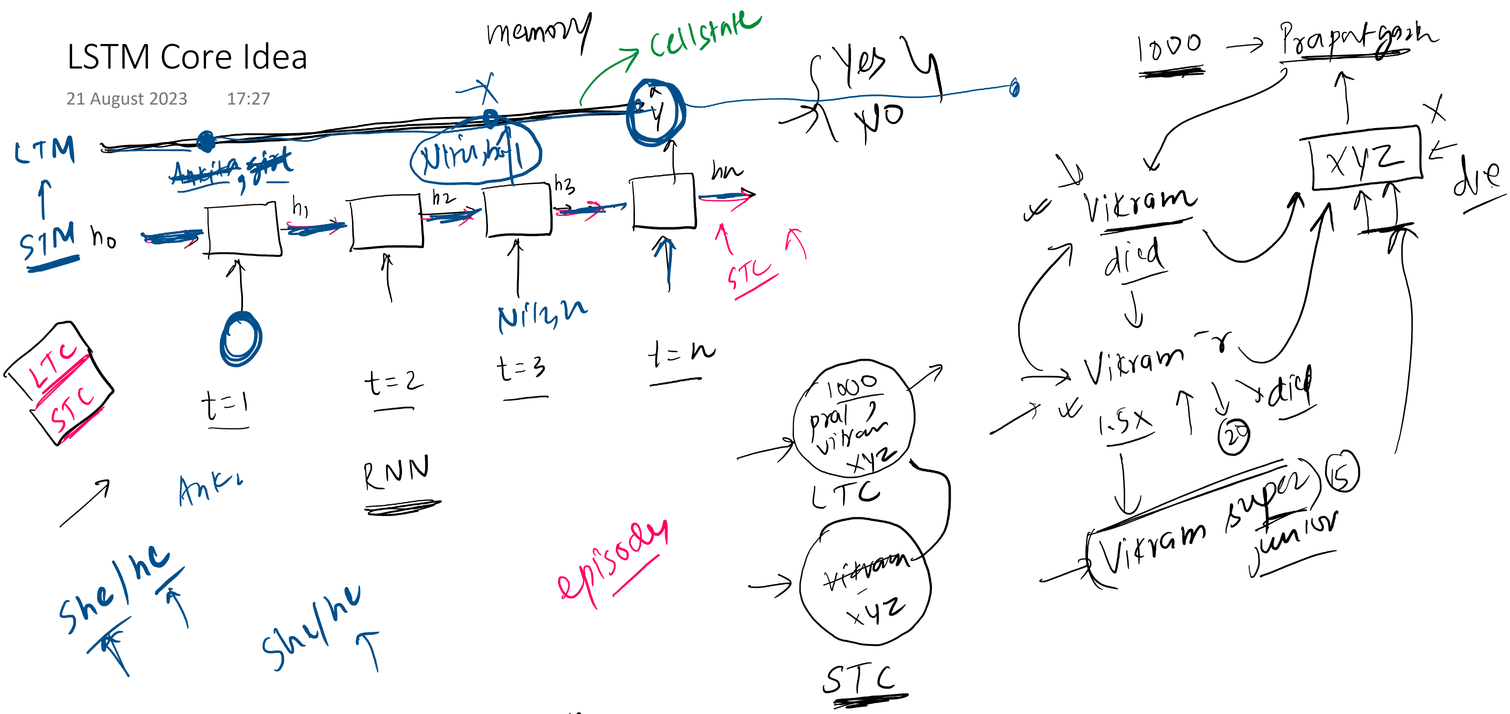
Recap

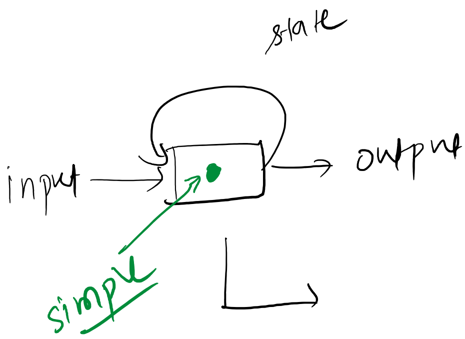
21 August 2023 11:55

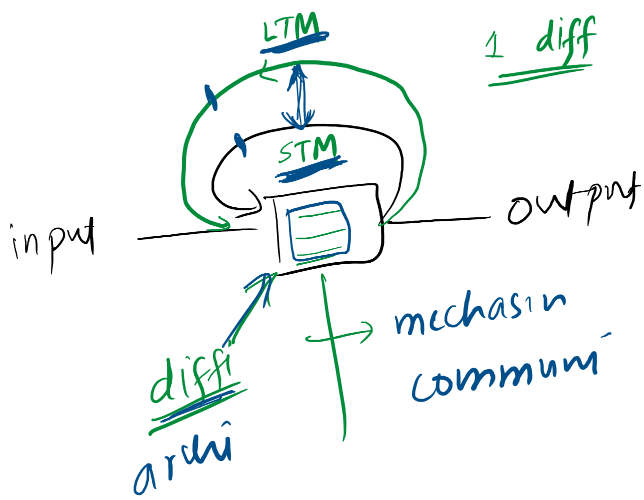


LSTM Core Idea

21 August 2023 17:27

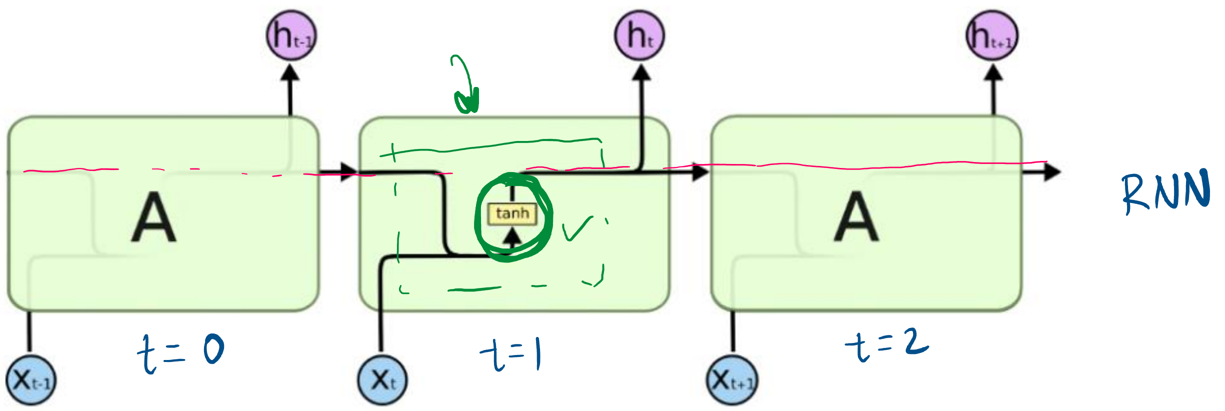


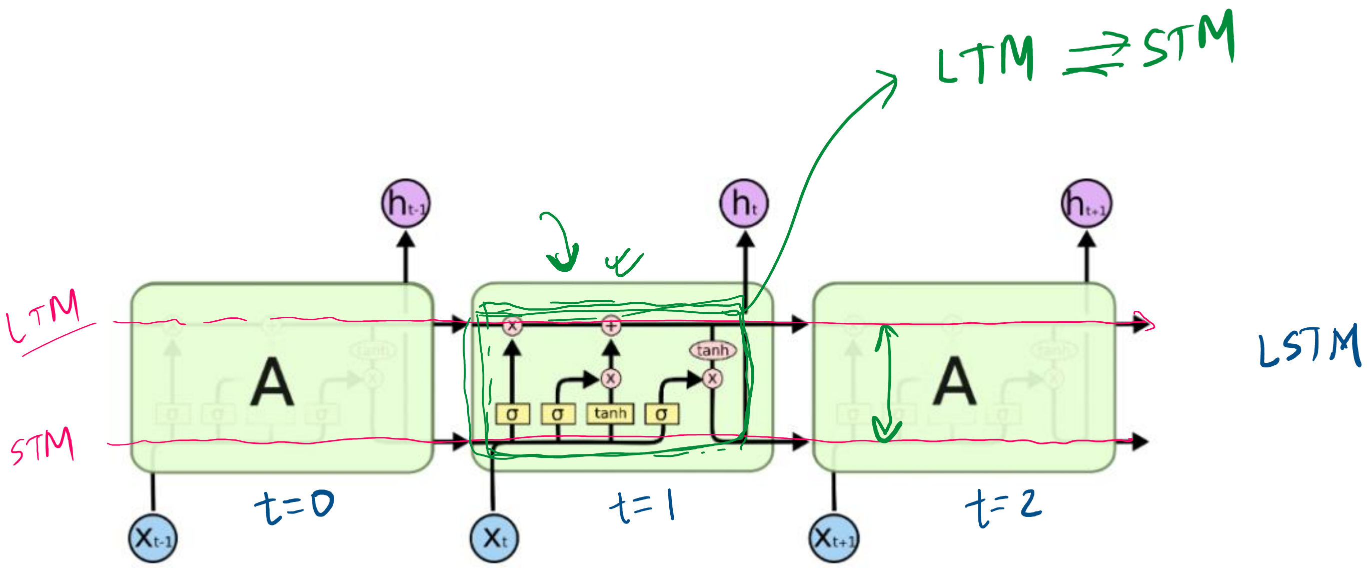


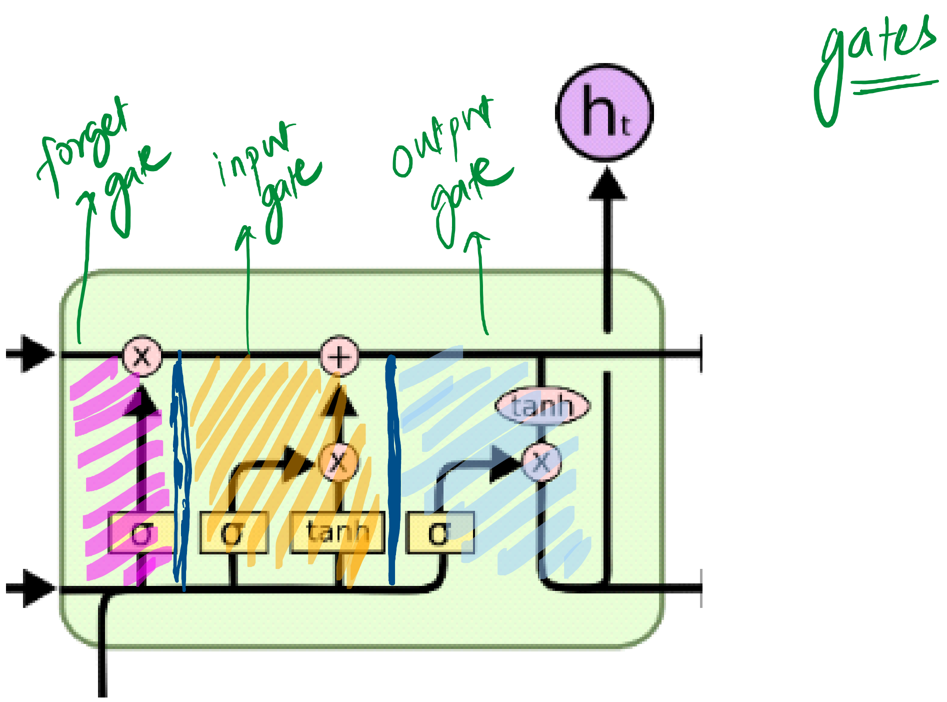


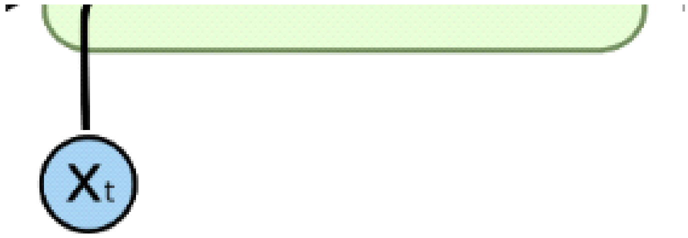
LSTM Architecture

21 August 2023 18:41



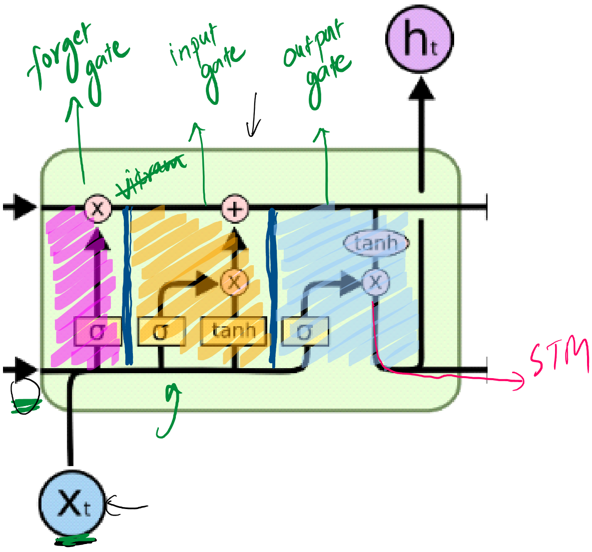





Xt

LSTM Gates

21 August 2023 19:06



forget

input

LTM

remove

1000 prarthna vikram jr

yes

No

input

output

output gate

Short term memory

LTM → add

create

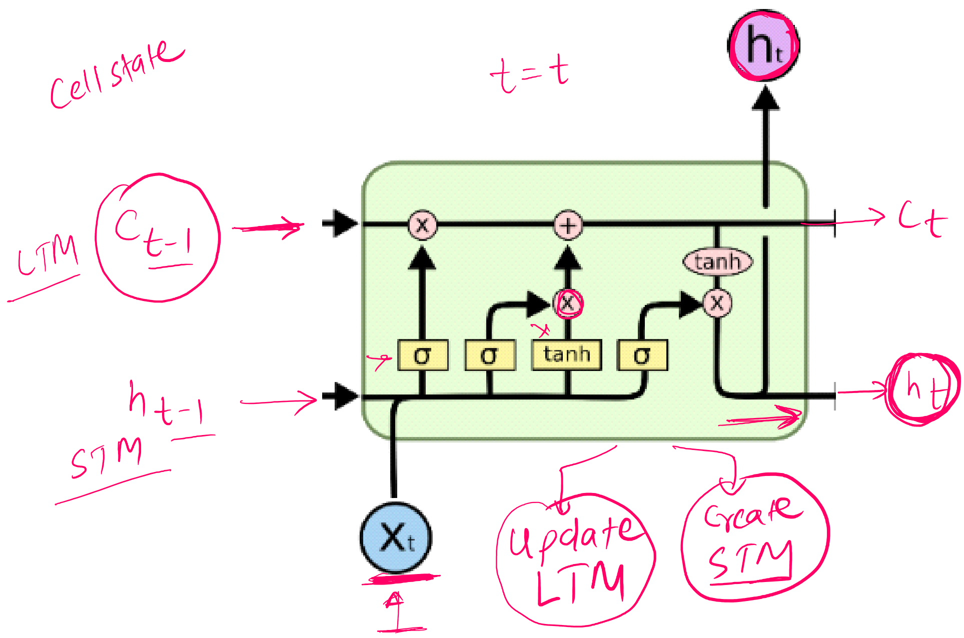
Summary

21 August 2023 19:29

LSTM

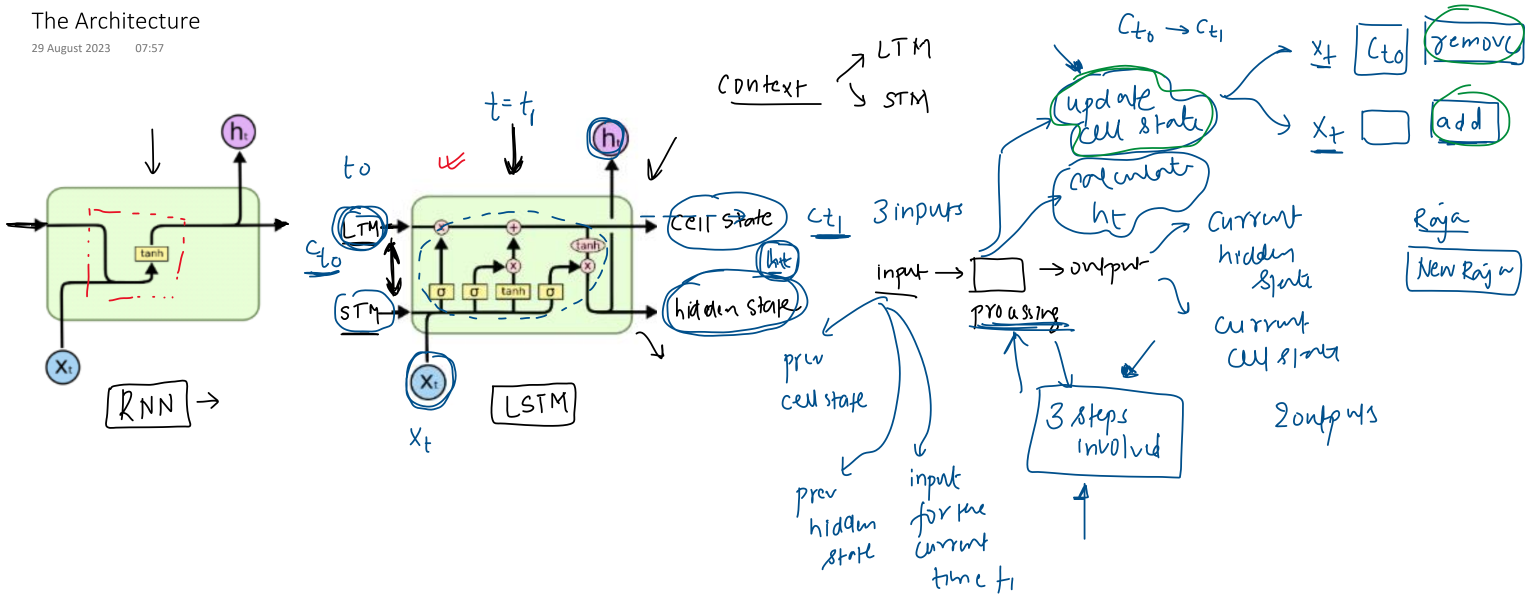
input → □ → output

processing



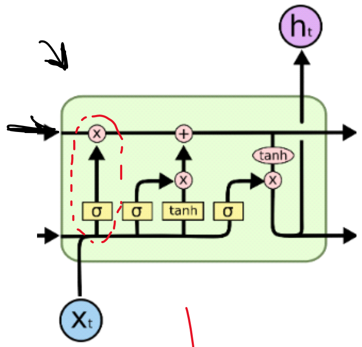
The Architecture

29 August 2023 07:57

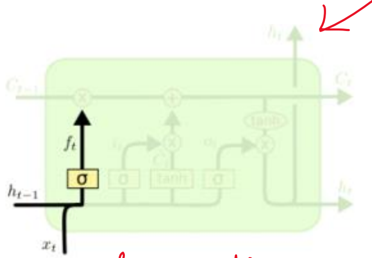


The Gates

29 August 2023 08:09



3 gates

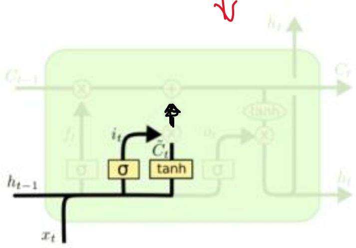


update ct

forget gate

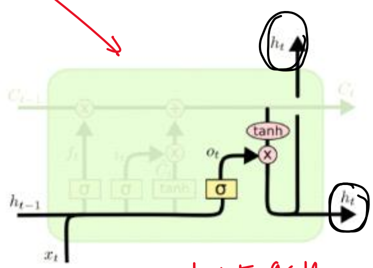
to remove something from Ct

+



input gate

to add somethin to the Ct

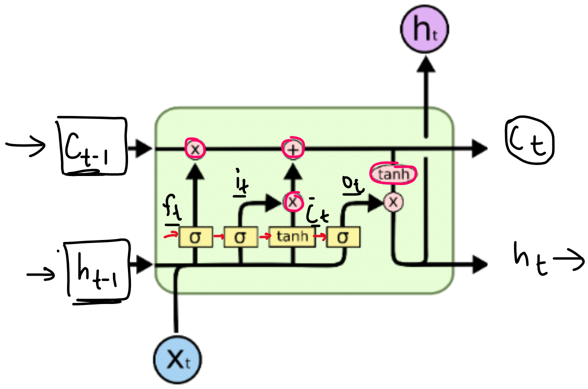


output gate

calculat hidden state ht

What are C_t and h_t

29 August 2023 08:08



$\boxed{h_t \quad c_t}$ → vectors

→ [0.1 0.3 0.9]

3d vector

rule

$\boxed{\text{same}}$

$\boxed{h_t \quad c_t}$ dim equal

h_t [0.1 0.45 0.6]

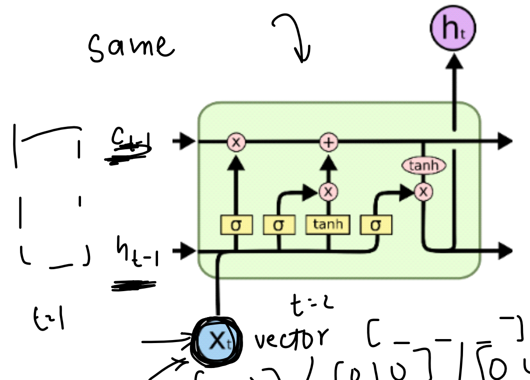
c_t [0.55 0.6 0.0]

What is X_t

29 August 2023 17:40

RNN

same



vector [_ _ _]

[1 0 0] / [0 1 0] / [0 0 1]

t=1 t=2 t=3

sentiment

vectorization

cat mat rat	0
cat rat rat	0
mat mat cat	1

[1 0 0] [0 1 0] [0 0 1]

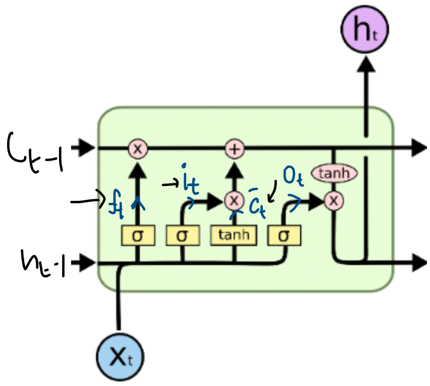
word vector

OHE

cat	mat	rat
1	0	0
0	1	0
0	0	1

What are f_t, i_t, o_t and $\bar{C}t$

29 August 2023 08:09



$\bar{C}_t$ → candidate cell state

$\underbrace{f_t}_{\text{forget gate}}$ $\underbrace{i_t \quad \bar{C}_t}_{\text{input gate}}$ $\underbrace{o_t}_{\text{output gate}}$

vectors → $\boxed{C_t \ h_t}$

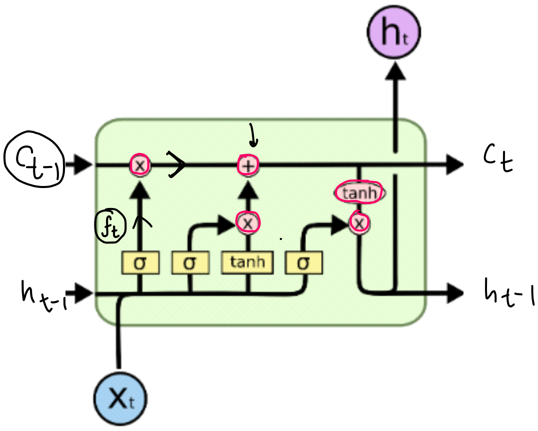
f_t i_t $\bar{C}_t$ o_t

[x y z]

[ ]

Pointwise Operations

29 August 2023 18:26



→ ⊗

→ ⊕

→ tanh

$C_{t-1} = [4 \ 5 \ 6] \rightarrow [0.26 \ 0.34 \ 0.53]$

tanh(4)

$f_t = [1 \ 2 \ 3]$

$[5 \ 7 \ 9]$

shape/dim

↓

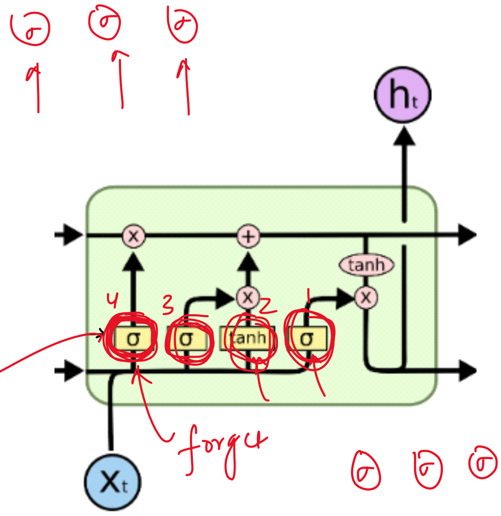
vector

$C_{t-1} \otimes f_t \rightarrow$ vector

$\rightarrow [4 \ 10 \ 18]$

→Neural Network Layers

29 August 2023 18:34



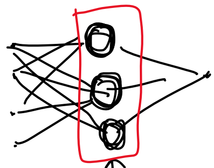
num units nodes

forget

hyperparameter

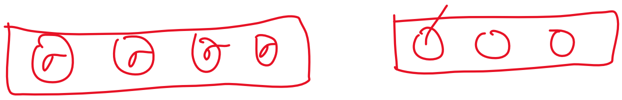
↳ (4)

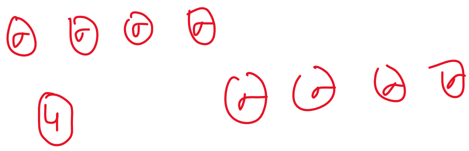
ANN



hidden

tanh





(4)

The Forget Gate

29 August 2023 19:58

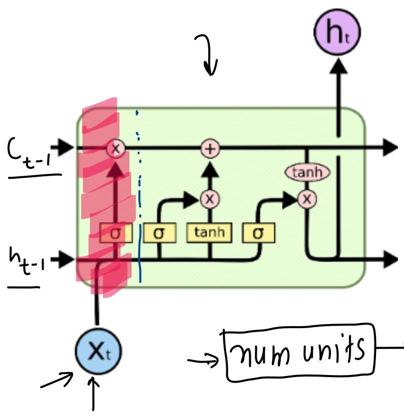
vectors dim

③ dim

input
output → $\boxed{\text{cell state}}$ → $\boxed{\text{remove}}$

→ 1) calculate f_t

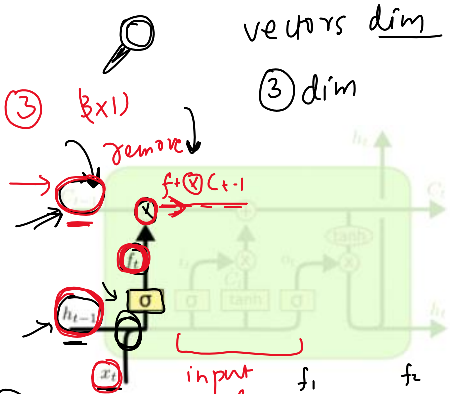
2) $\boxed{C_{t-1} \otimes f_t}$



→ $\boxed{\text{num units}}$ → ③

4 dim vector

[x_{i1} x_{i2} x_{i3} x_{i4}]



③ (3x1) remove

$f_t \otimes C_{t-1}$

input gate (add)

$(3\times1) \otimes (3\times1)$

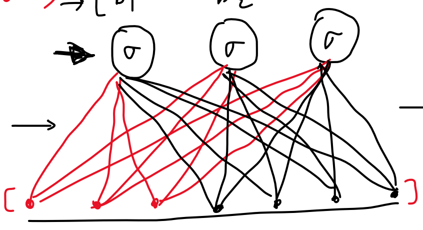
σ σ σ

$f_3 \rightarrow [f_1\ f_2\ f_3] \rightarrow \boxed{f_t\ (3)}$

f_1 | b_1 f_2 | b_2 f_3 | b_3 (3x1) (3 x 7) (7x1)

→ (3x1) + (3x1) = (3x1)

[ ] →



→ $\boxed{W_{3\times7}}$ (21 weights) biases

21 wts
4 biases

h_{t-1} (3) x_t (4) → 7x1

forget gate ← 0%
50%
100%

remove

$[h_{t-1}, x_t]$

long term

f_t

$$\underbrace{f_t}_{(3\times1)} = \sigma\left(\underbrace{\underbrace{W_f}_{(3\times7)}\underbrace{[h_{t-1}, x_t]}_{(7\times1)}}_{(3\times1)} + \underbrace{b_f}_{(3\times1)}\right) \rightarrow (3\times1) \rightarrow (3\times1)$$

[4 5 6]

C_{t-1} → ⊗ → [0 0 0] ⊗

f_t → [$\frac{1}{2}$ $\frac{1}{2}$ $\frac{1}{2}$]

[1 1 1]

[2 2.5 3]

[4 5 6]

[0 0 0] → 100% remove

removal

$\boxed{f_t \otimes C_{t-1}}$ → (3x1)

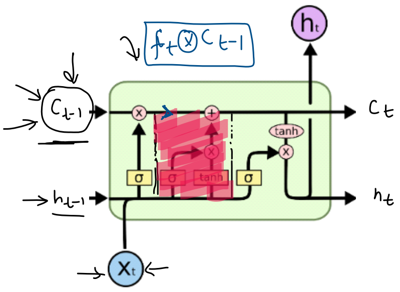
(C_{t-1}) → remove

The Input Gate

30 August 2023 04:38

add some new imp info to the C_t

$f_t \otimes C_{t-1}$



stages

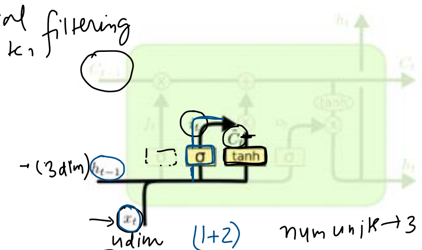
1) $\boxed{\bar{C}_t}$ candidate cell state

2) i_t

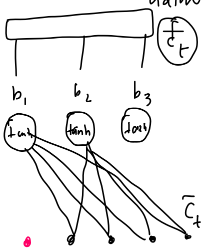
3) $\boxed{C_t}$ → current cell state

potential important info

filtering



$\boxed{\bar{C}_t} = (3 \times 1)$



→ W_c (3×7)

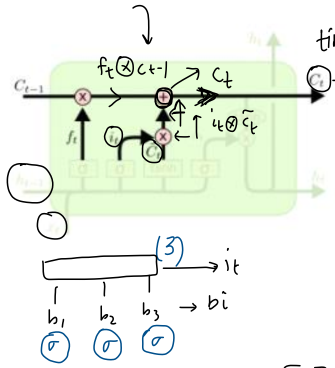
$\bar{C}_t = \tanh(W_c[h_{t-1}, x_t] + b_c)$

h_{t-1} x_t

(7×7) (7×1) (3×1)

(3×1)

(3×1) ↓ (3×1)



→ [ ] W_i

h_{t-1} x_t

$i_t = \sigma(W_i[h_{t-1}, x_t] + b_i)$

(3×7) (7×1) (3×1)

(3×1)

σ(3×1) → (3×1)

pointwise

$i_t \otimes \bar{C}_t \rightarrow \bar{C}_t^*$ (filtered candidate cell state)

(3×1) (3×1)

[0.5 0.5 0.5] [4 5 6]

1 1 1 ↓ 50% 0% 100%

[2, 2.5 3]

$\boxed{C_t = f_t \otimes C_{t-1} \oplus i_t \otimes \bar{C}_t}$

i_t - filter

$\bar{C}_t \rightarrow C_t$

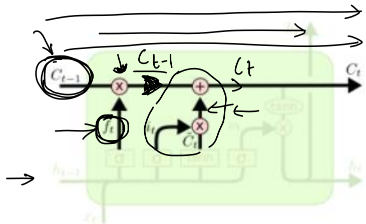
10%

[4 5 6] → [4 5 6]

t = t-1   t = t

$f_t = [1\ 1\ 1]$

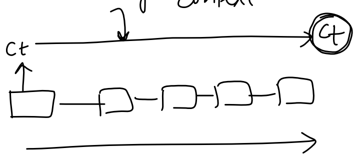
$i_t \otimes \bar{C}_t = [0\ 0\ 0]$

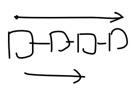


[4 5 6] [0 0 0]

$C_t = [4\ 5\ 6]$

long term context





The Output Gate

30 August 2023 05:27

current timestep hidden state

$C_t \to h_t$

make

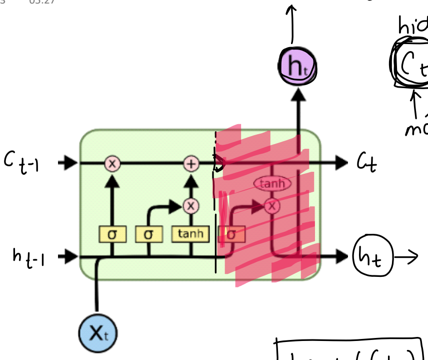
$\tanh(C_t)$ $[-1\ 1]$

filtration $\to \tanh(C_t)$

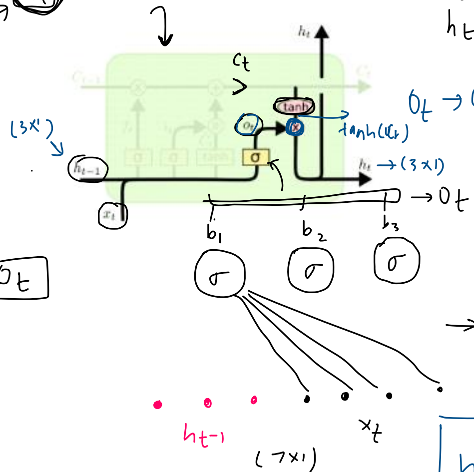
$\to o_t$

$h_t = [o_t \otimes \tanh(C_t)]$

$\tanh(C_t) \to (3\times1)$



$\boxed{\tanh(C_t)}$ $\boxed{o_t}$



$o_t \to (3\times1)$

$$\boxed{o_t = \sigma\left(\underbrace{W_o}_{3\times7}\underbrace{[h_{t-1}, x_t]}_{(7\times1)} + \underbrace{b_o}_{(3\times1)}\right)}$$

$(3\times1) + (3\times1)$

$\sigma(3\times1) \to (3\times1)$

$\to W_o [\ \ ]_{3\times7}$

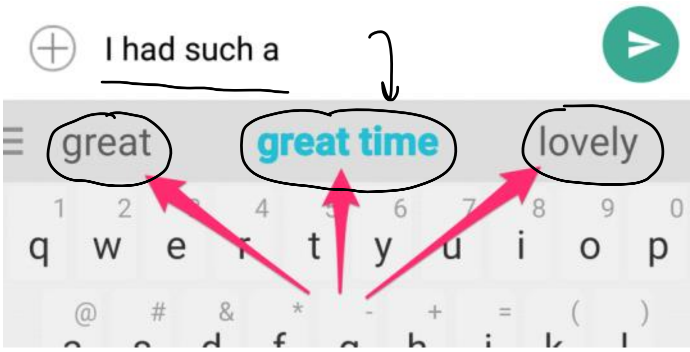
$$\boxed{h_t = \underbrace{o_t}_{(3\times1)} \otimes \underbrace{\tanh(C_t)}_{(3\times1)} \to (3\times1)}$$

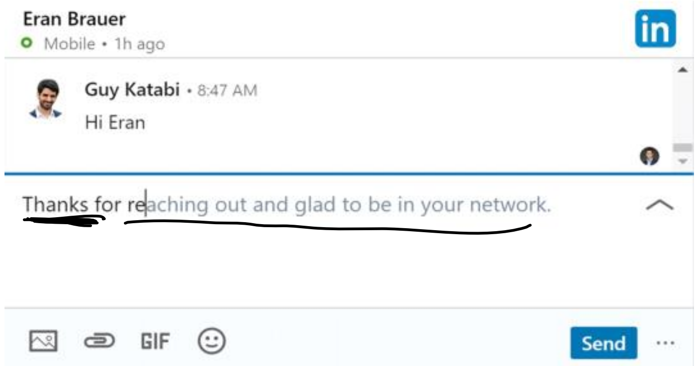
output current time step

What is a Next Word Predictor

08 September 2023 08:50

code





The Strategy

08 September 2023 08:51

[Text generator] ——

→ supervised learning

↓

input | output

model → input → output

9 words

1	2	3	4	5
Hi	my	name	is	Nitish
I	live	in	gurgoan	
6	7	8	9	

numbers [001000000]

2.1 2.7

[010000000] → 2word

reg → discrete

LSTM → input | output

english

number input | output

dataset

input	output
Hi	my
Hi my	name
Hi my name	is
Hi my name is	Nitish
I	live
I live	in
I live in	gurgoan

input	output
[1]	2
[1,2]	3
[1,2,3]	4
[1,2,3,4]	5
[6]	7
[6,7]	8

→ OHE

LSTM

classification → binary, multiclass

about the program

[93 1 13]

input	out
[93]	1
[93,1]	13

text gen → supervised → Reg, Classification

~~9 wo~~ 3 words 1

↓

→ [0.62 0.51 0.31]

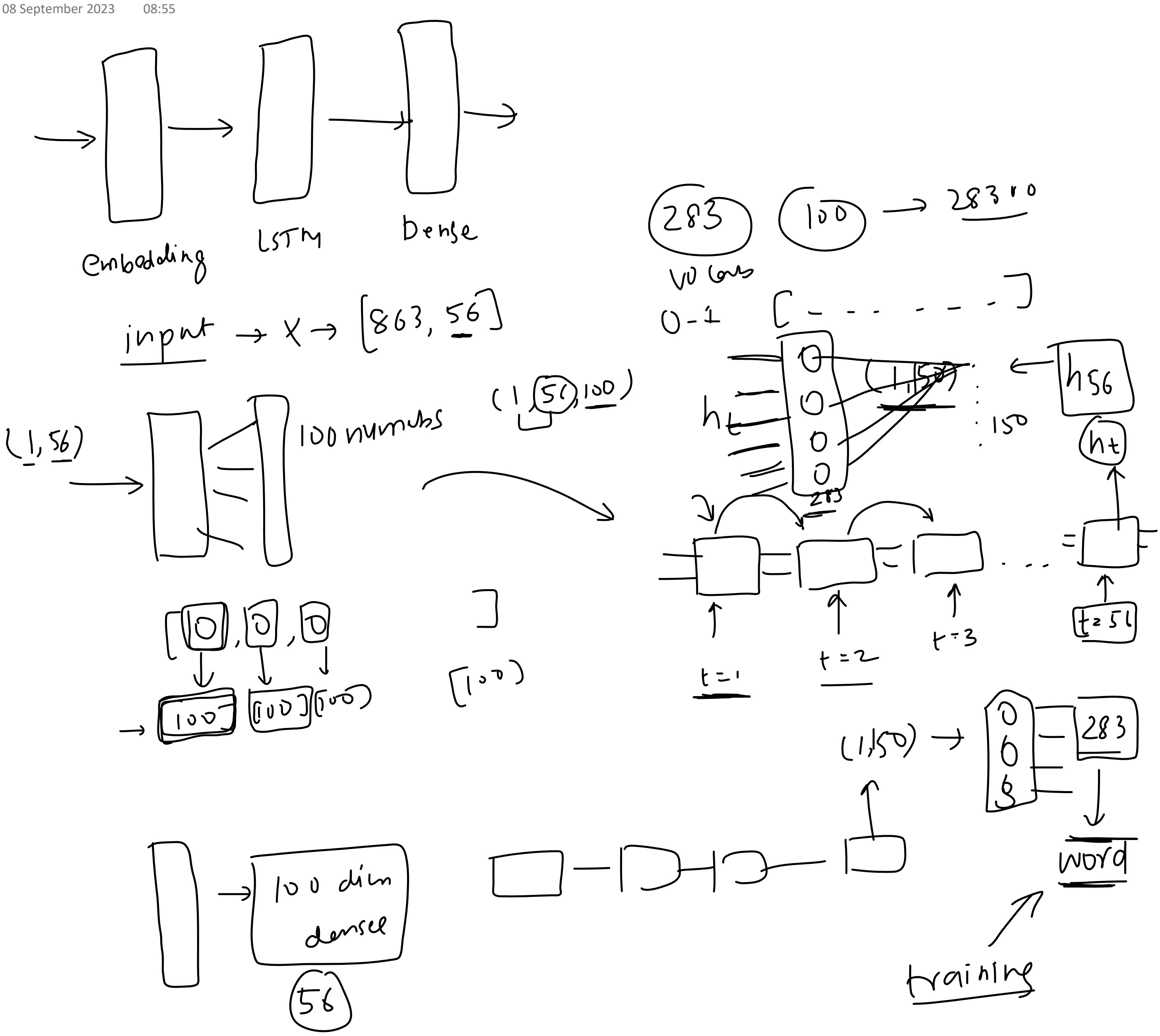
[0.33 0.41 0.21]

[1 0 0]
↳ first word

[0 1 0]
↳ second word

The Architecture

08 September 2023 08:55



How to improve performance?

08 September 2023 08:51

1> more data → $\boxed{283}$

2> Hyperparameter →

3> Advanced arch

- Stacked LSTM
- Bidirection LSTM
- Transformers

What is GRU

05 October 2023 00:09

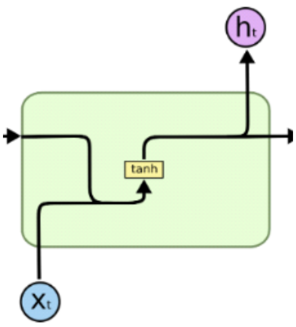
sequence → long term

3 gates

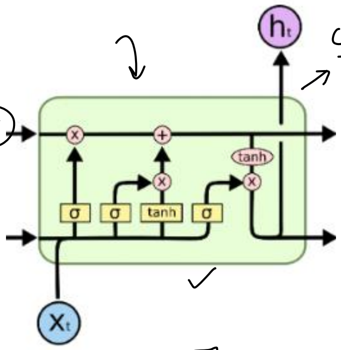
RNN archi

complex parameters → training more

2 gates



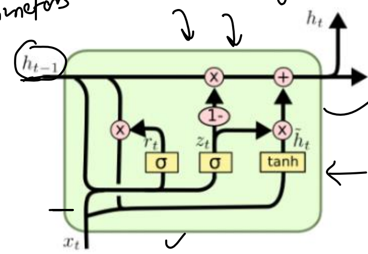
Simple RNN



→ LSTM ≪

(97)

⟷



less parameters

simpler

≪ GRU →

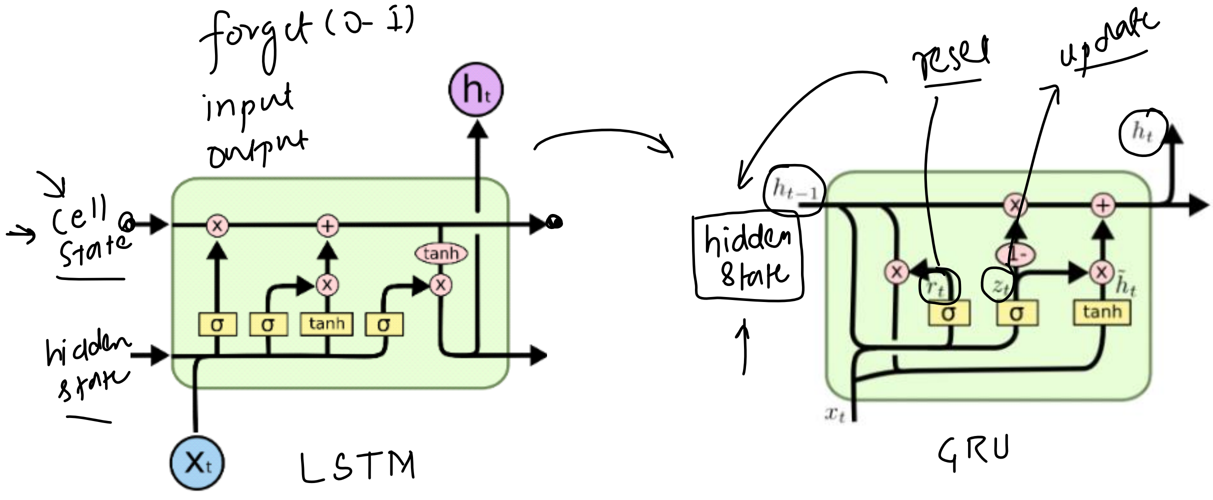
2014

≃ LSTM

1) Archi
2) Example
3) lstm vs gru

The Big Idea Behind GRU

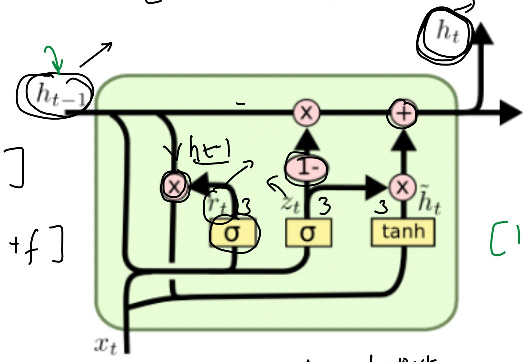
05 October 2023 00:47



The Setup

05 October 2023 01:07

[a b c] [d e f]



[ad be cf]

[a+d b+e c+f]

neural network layers (FC layers)

→ Advise → LSTM/GRU → confusing

goal → $\boxed{t}$

h_{t-1}, x_t → h_t

prev hidden state, curr hidden state, input

h_{t-1}, h_t, $\boxed{x_t}$, r_t, z_t, $\tilde{h}_t$ → vectors

? → candidate hidden state

reset gate

update gate

[1 2 3 4]

→ $\boxed{\sigma}$ $\boxed{tanh}$?

× + 1- ?

[1 2 3 4] 4dim same

number of = 5 6

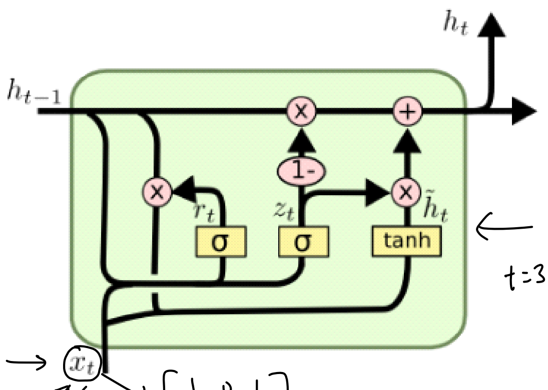
10

$\boxed{\sigma}$ → σ σ σ σ σ

The Input Xt

05 October 2023 01:52

t input → vector



[1 0 1]

sentiment

$\boxed{\text{OHE / BOW / word2vec}}$

text	review
cat mat rat	1
cat rat rat	0
mat rat mat	1

t=3

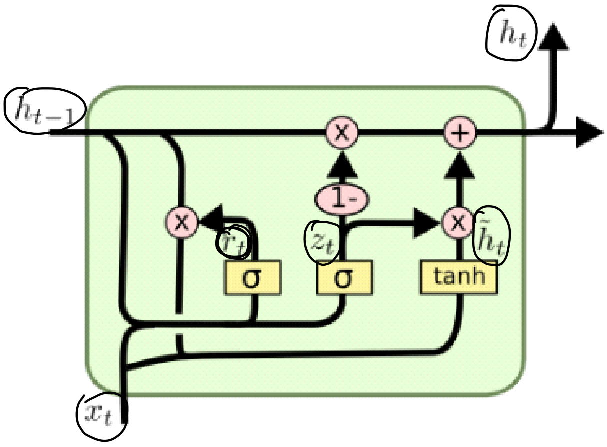
S1 [1 0 0] [0 1 0] [0 0 1]

S2 [1 0 0] [0 0 1] [0 0 1]

S3 [0 1 0] [0 0 1] [0 1 0]

Architecture

05 October 2023 02:10



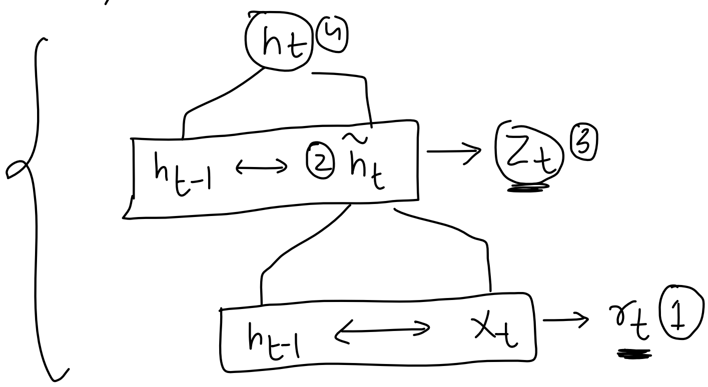
$h_{t-1}, x_t \longrightarrow h_t$

→ 1) Calculate r_t (reset gate)

→ 2) Calculate $\tilde{h}_t$ (candidate hidden state)

→ 3) Calculate z_t (update gate)

→ 4) Calculate h_t (current hidden state)



What exactly is hidden state?

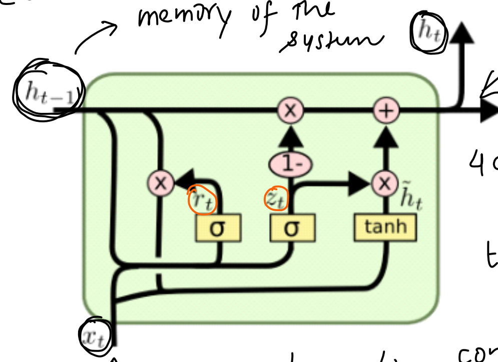
05 October 2023 02:19

t → h_{t-1}, x_t → h_t → vector [○ ○ ○ ○]

memory → context

[0 0 0 0]

memory of the system



t=2

t=3

t=4

4 dim

t=1

→ happy/sad

- [There was a king Vikram very strong and powerful]
- There was an enemy king kaali
- Both had a war and kaali killed Vikram
- Vikram had a son Vikram Jr who grew up he to become very strong just like his father
- He also attacked Kaali But got killed
- Vikram Jr too had a son called Vikram super Jr and when he grew up he also fought kaali
- And he killed kaali and took revenge of his father and grand father

h_{t-1}

h_{t-1} / h_t context

→ [0.5 0.6 0 3 0.1]

→ [Power, Conflict, Tragedy, Revenge]

→ [0.9 0.6 0.9 0.9]

→ [0.9 0 0 0] h_t

→ [1 0.5 0 0] → h_{t-1}

[0.7 0.7 0.4 0]

[0.9 0.6 0.3 0]

[0.8 0.8 0.6 0.5]

[1 0.9 0.5 0.7]

[0.7 0.8 0.3 1]

$h_{t-1} \rightarrow h_t$ GRU

↓ 2 steps gates

reset update

① $h_{t-1} \rightarrow \tilde{h}_t$ ② → h_t

candidate memory

reset gate r_t

update gate z_t

[power, conflict, tragedy, revenge]

→ [0.5, 0.6, 0.7, 0.1] → h_{t-1}

x_t → [0.8, 0.4, 0.5, 0.3] → $\tilde{h}_t$ candidate

inclined x_t

h_t

$\tilde{h}_t$ memory → candidate → h_t

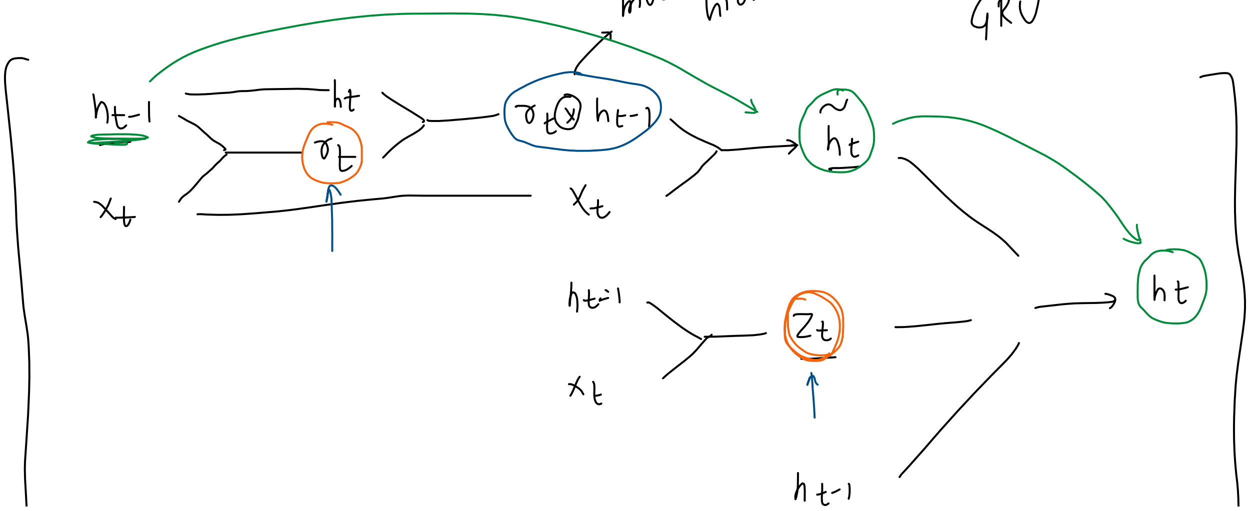
h_{t-1} $\tilde{h}_t$ → h_t

[0.65, 0.5, 0.6, 0.2] → h_t

$h_{t-1} \xrightarrow{r_t} \tilde{h}_t \xrightarrow{z_t} h_t$

modulated/resetted hidden state

GRU



h_{t-1}

Calculating the reset gate

05 October 2023 14:24

$\sigma\left(\boxed{\text{weighted sum}}\right)$

$\boxed{0 \quad 1}$

h_{t-1} — h_t, r_t → $r_t \otimes h_{t-1}$, x_t → h_t

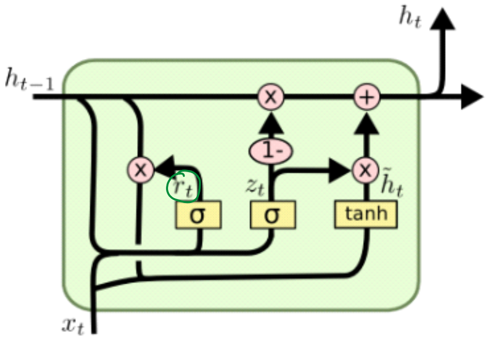
[? , ? , t , r]

[0 1 _ _ _]

$\boxed{\text{reset}}$

r_t → vector → h_{t-1}

↳ gate



lower lower higher

$h_{t-1} = [0.6, 0.6, 0.7, 0.1]$

$r_t = [0.8, 0.2, 0.1, 0.9]$

gate

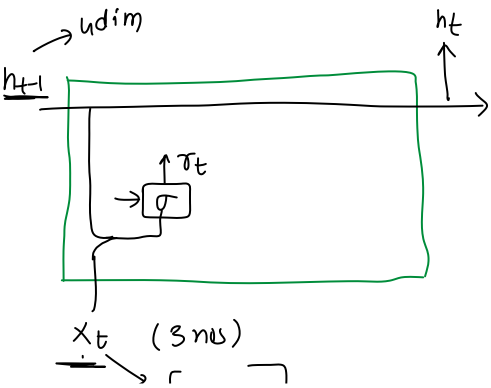
80% percent 20% 90%

h_{t-1}, x_t → z_t [0 - 1] → h_t

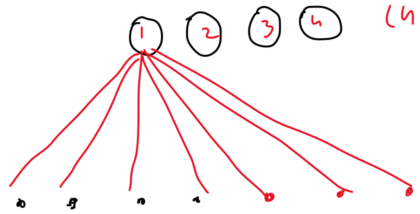
[{ vikram, kaki, war }] h_{t-1}

reset

h_t → vikram Jr x_t



$$r_t = \sigma\left(W_r[h_{t-1}, x_t] + b_r\right)$$



(4)

← (7×4) = $\boxed{28 + 4}$

$(7) \times 1 \rightarrow (1 \times 7)$

h_{t-1}

modulated resetted

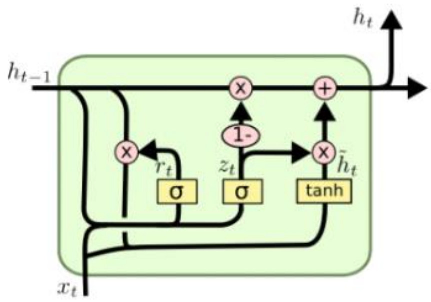
$(1 \times 7)\ (7 \times 4)$

$(1 \times 4) + (1 \times 4) \rightarrow ((1 \times 4))$

h_{t-1}, x_t → $\boxed{h_{t-1}}$, $\boxed{r_t}$ → $\boxed{r_t \otimes h_{t-1}}$, x_t → $\tilde{h}_t$

h_t

$(1 \times 4) \rightarrow r_t$



→ $h_{t-1} = [0.6, 0.6, 0.7, 0.1]$

→ $r_t = [0.8, 0.2, 0.1, 0.9]$

$\boxed{\tilde{h}_t} = [0.7\ 0.2\ 0.1\ 0.2]$

↑ x_t

90%

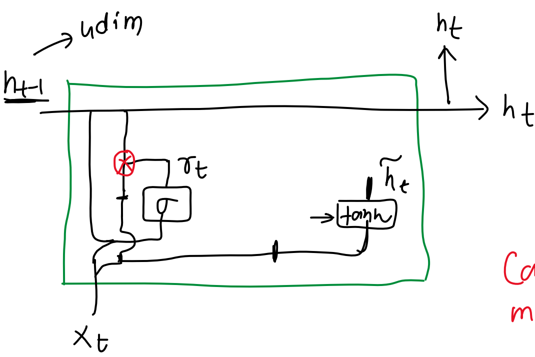
power conflict tragedy revenger

→ $\boxed{h_{t-1} \otimes r_t} = [0.48, 0.12, 0.07, 0.09]$ → x_t

h_{t-1} / $\tilde{h}_t$

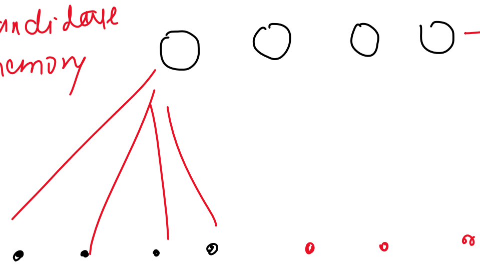
↓

z_t



$$\rightarrow \boxed{\tilde{h}_t} = \tanh\left(W_c[h_{t-1} \otimes r_t, x_t] + b_c\right)$$

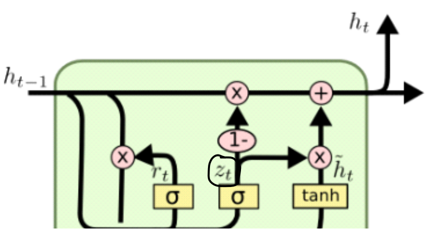
Candidate memory



→ b_c

→ W_c (7×4)

$(h_{t-1} \otimes r_t)$

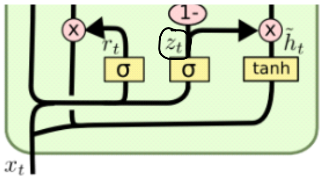


h_{t-1}, x_t → h_{t-1}, r_t → $r_t \otimes h_{t-1}$, x_t → $\tilde{h}_t$

$\tilde{h}_t = [0.7\ 0.2\ 0.1\ 0.2]$

$h_{t-1} = [0.6\ 0.6\ 0.7\ 0.1]$

h_{t-1} → z_t



h_{t-1} — L

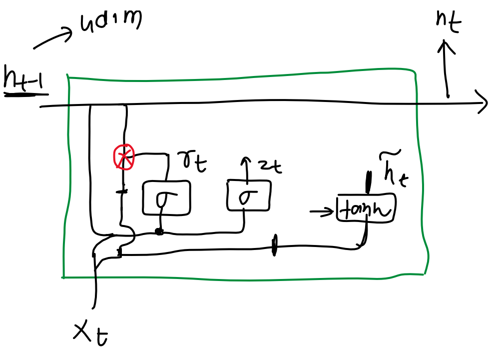
[[— — — —]]

(0–1)

h_{t-1}, x_t → z_t

x_t → important

x_t → not imp



gate

h_{t-1} and $\tilde{h}_t$ (h_t)

$$z_t = \sigma(W_z[h_{t-1}, x_t] + b_z)$$

(1x4) h_{t-1}

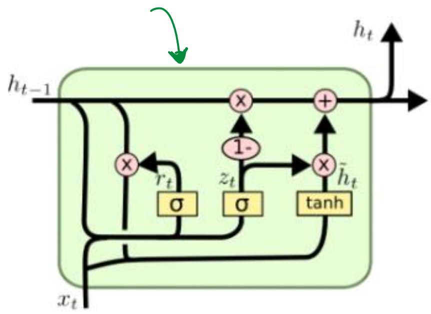
$\tilde{h}_t$

h_t

○ ○ ○ ○ → bias(4)

→ (7x4)

h_{t-1}



$\tilde{h}_t$ = [0.7 0.2 0.1 0.2]

h_{t-1} = [0.6 0.6 0.7 0.1]

z_t = [0.1 0.7 0.8 0.2]

0.1 0.1

h_t [0.07 0.14 0.08 0.04]

$$h_t = (1 - z_t) \otimes h_{t-1} \oplus z_t \otimes \tilde{h}_t$$

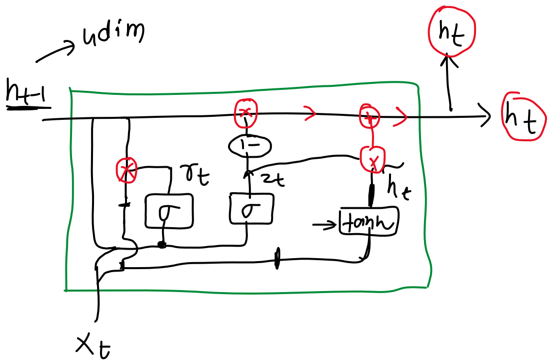
$(1-z_t) \to 0.9$

$(1-z_t)$ [0.9 0.3 0.2 0.8] [0.6 0.6 0.7 0.1]

[0.54 0.18 0.14 0.08]

→ h_t [0.61 0.32 0.22 0.12]

power conflict tragedy revenge



h_{t-1}, x_t → h_{t-1}, r_t → $r_t \otimes h_{t-1}$, x_t → $\tilde{h}_t$

h_{t-1}, x_t → z_t

h_{t-1}, z_t, $\tilde{h}_t$ → h_t

LSTM vs GRU

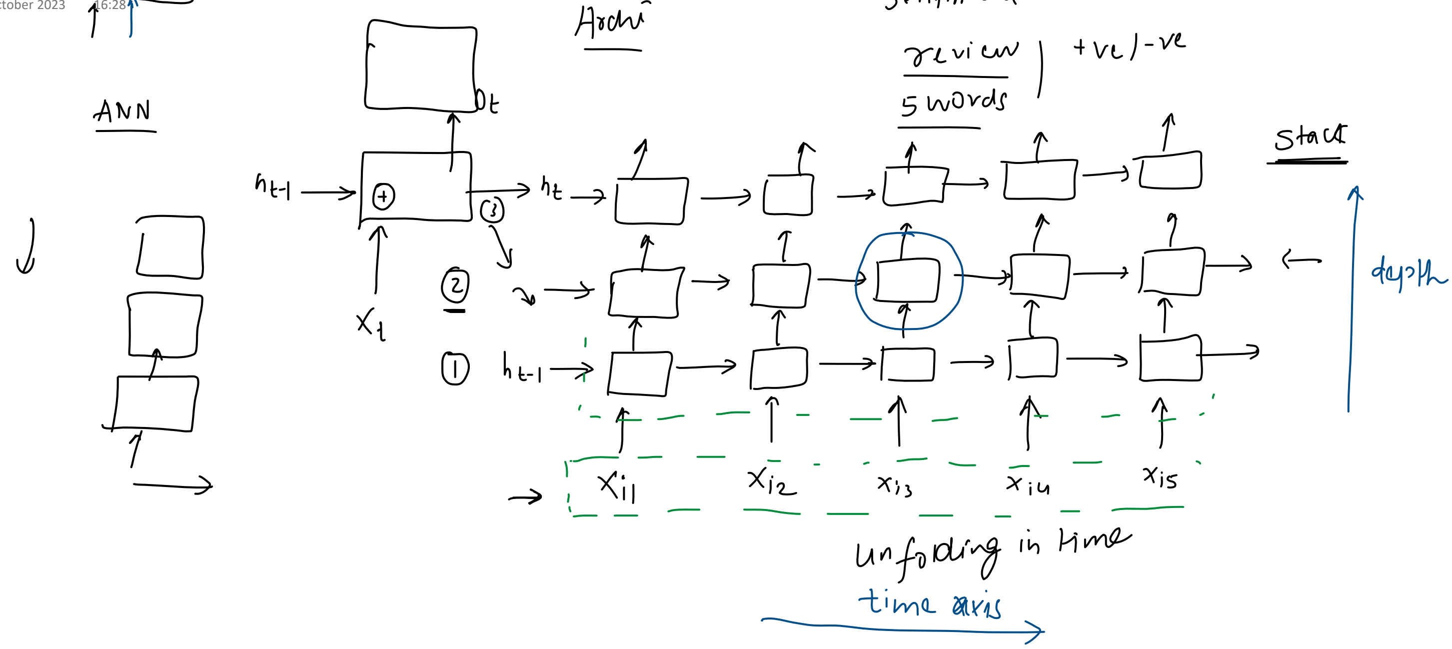
05 October 2023 16:45

Here are the main differences between LSTM and GRU:

1. Number of Gates:
   - LSTM: Has three gates — input (or update) gate, forget gate, and output gate.
   - GRU: Has two gates — reset gate and update gate.
2. Memory Units:
   - LSTM: Uses two separate states - the cell state (ct) and the hidden state (ht). The cell state acts as an "internal memory" and is crucial for carrying long-term dependencies.
   - GRU: Simplifies this by using a single hidden state (ht) to both capture and output the memory.
3. Parameter Count:
   - LSTM: Generally has more parameters than a GRU because of its additional gate and separate cell state. For an input size of d and a hidden size of h, the LSTM has $4 \times ((d \times h)+(h \times h)+h)$4×((d×h)+(h×h)+h)parameters.
   - GRU: Has fewer parameters. For the same sizes, the GRU has $3 \times ((d \times h)+(h \times h)+h)$3×((d×h)+(h×h)+h)parameters.
4. Computational Complexity:
   - LSTM: Due to the extra gate and cell state, LSTMs are typically more computationally intensive than GRUs.
   - GRU: Is simpler and can be faster to compute, especially on smaller datasets or when computational resources are limited.
5. Empirical Performance:
   - LSTM: In many tasks, especially more complex ones, LSTMs have been observed to perform slightly better than GRUs.
   - GRU: Can perform comparably to LSTMs on certain tasks, especially when data is limited or tasks are simpler. They can also train faster due to fewer parameters.
6. Choice in Practice:
   - The choice between LSTM and GRU often comes down to empirical testing. Depending on the dataset and task, one might outperform the other. However, GRUs, due to their simplicity, are often the first choice when starting out.

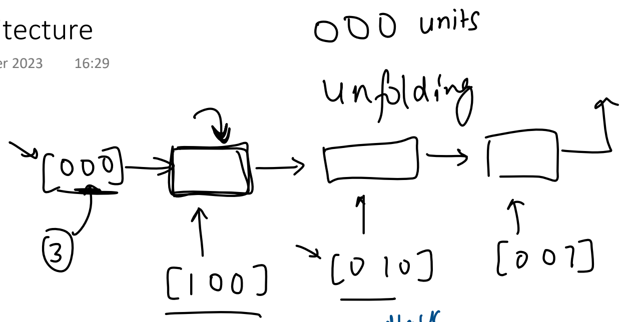
What is Deep RNN

17 October 2023 16:28



Architecture

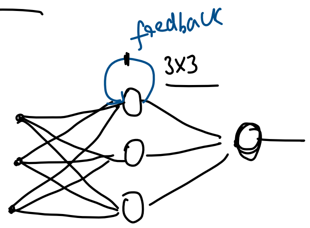
17 October 2023 16:29

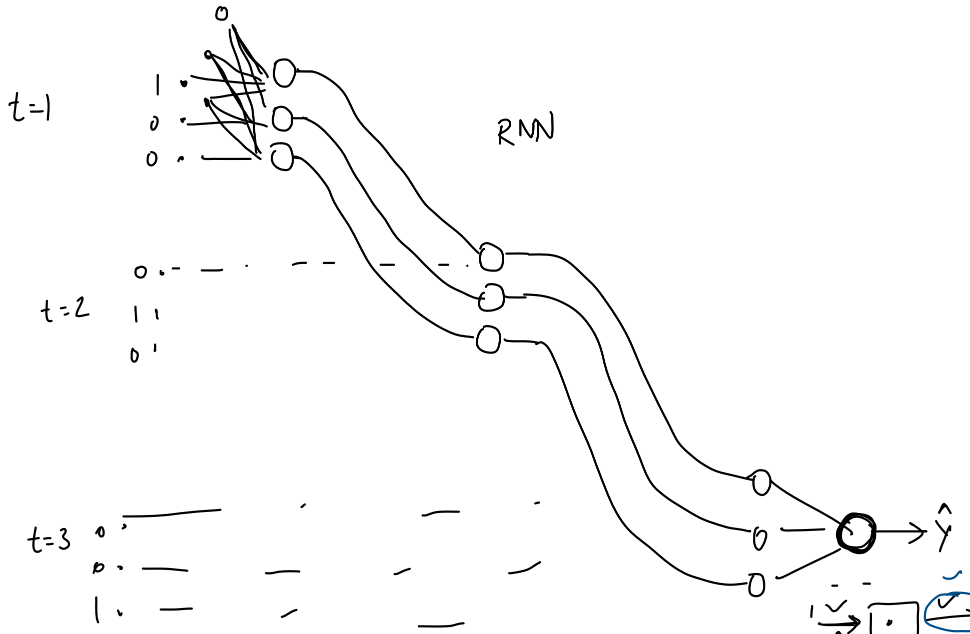


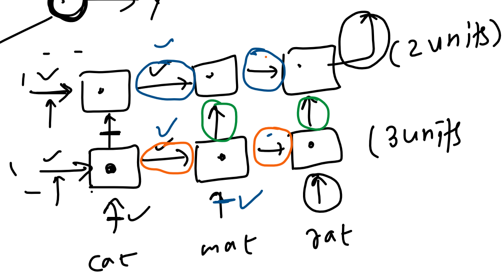
review	sentiment
cat mat rat	1
rat rat mat	0
mat cat cat	1

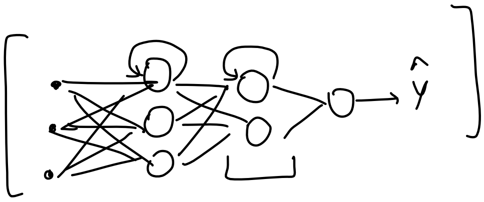
[1 0 0] [0 1 0] [0 0 1]

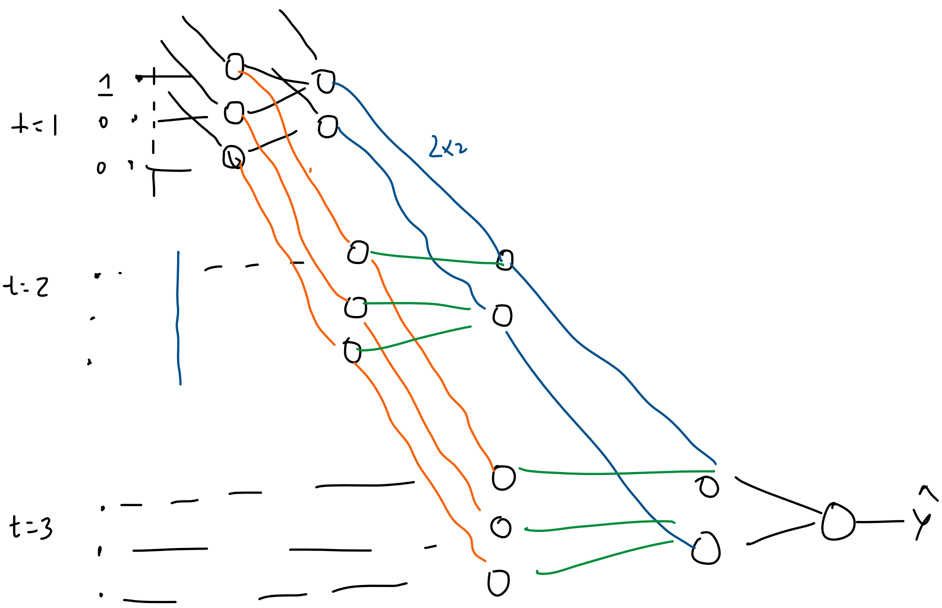
$\boxed{t=3}$ 3 words



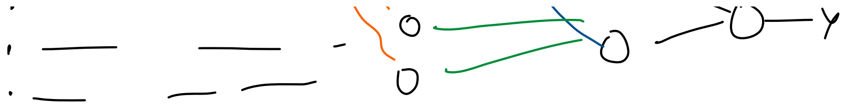






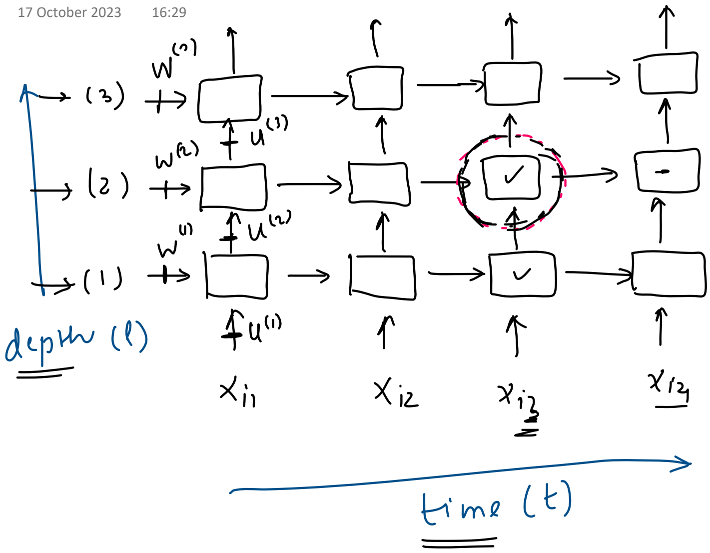


t=3

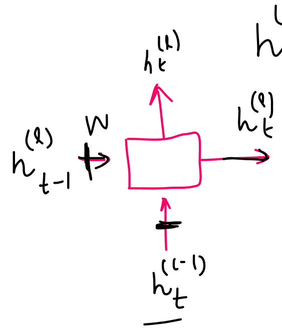


Notation

17 October 2023 16:29



$h_4^{(2)}$ $h_t^{(l)}$ $h_3^{(2)}$



$$h_t^{(l)} = \tanh($$

$$\left[h_t^{(l)} = \tanh\left(W^{(l)} h_{t-1}^{(l)} + U^{(l)} h_t^{(l-1)} + b^{l}\right) \right]$$

Why and When to use?

17 October 2023 16:29

1. Hierarchical Representation ✓
2. Customization for Advanced Tasks

deep RNN

stack

product

sentence

encoder-decoder

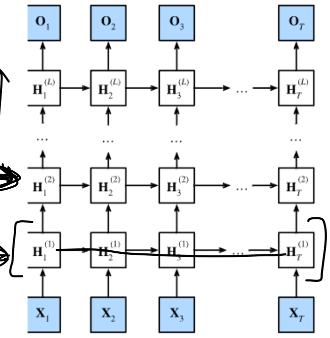
machine

deep RNNs

love hate amazing

terrible

deep RNN



→ primitive

Audio is bad

Display is great

Overall I am happy

→ sentiment

When to use Deep RNNs?

Complex tasks

Speech recog

Machine translation

Large datasets

Overfitting

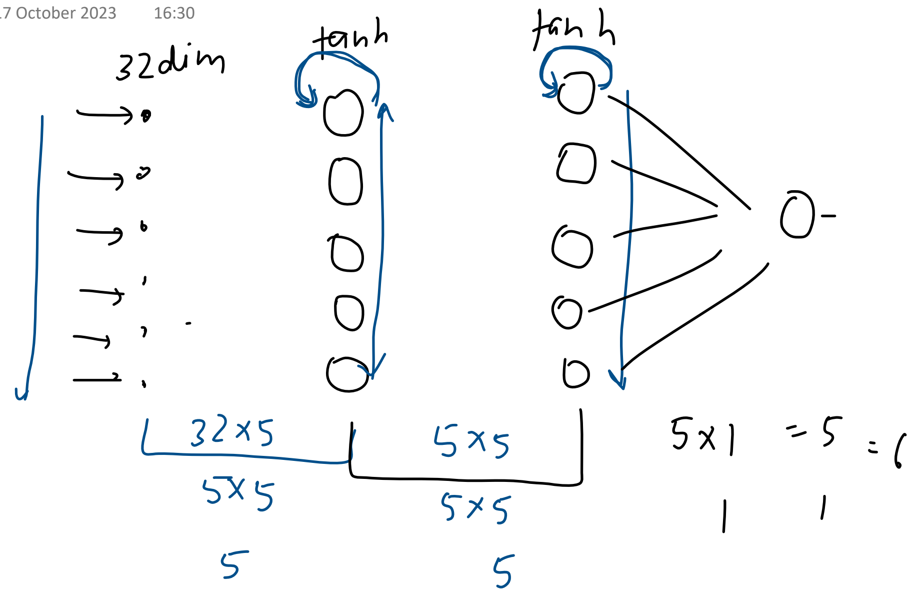
Computational

Simpler Models

Deep RNN

Code Example

17 October 2023 16:30



→ embedding

10000 × 32 = 3,20,000

SimpleRNN1

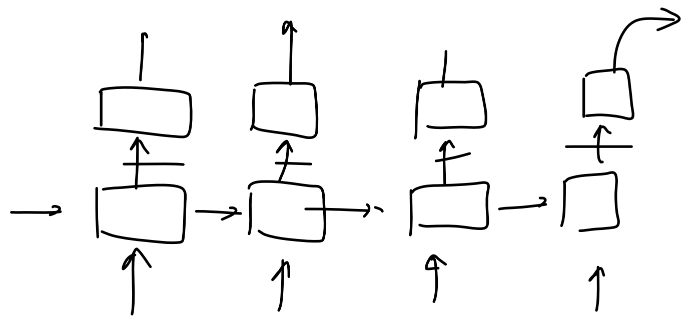
190

SimpleRNN2

55

Dense

6

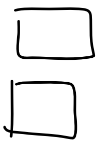


Variants

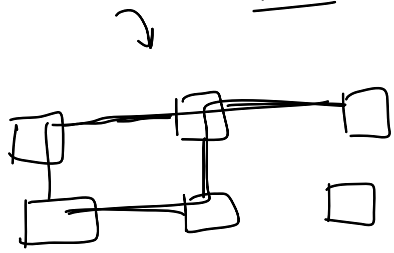
17 October 2023 16:30

SimpleRNN ✓✓

Deep LSTM ✓✓ Deep GRU ✓✓



Keras



[Computation graph]

Dropout

Disadvantages

17 October 2023 16:30

Overfitting

training time

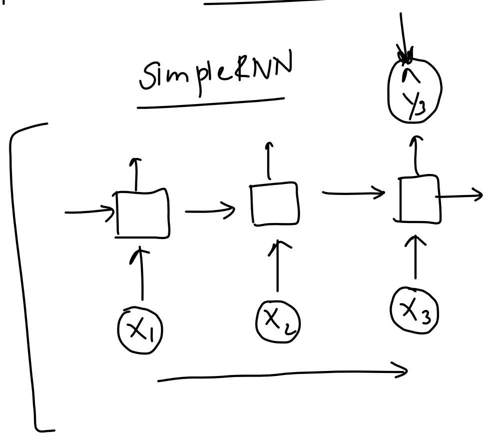
The Why? BiRNN

26 October 2023 15:19

RNN → LSTM, GRU → Deep RNN → BiRNN

unidirc RNN

SimpleRNN



[future inputs
past outputs

① [NER] → Named entity recog

NLP

Nilish → person
Gurgaon → loc/place } → chatbot

simpleRNN

ORG / LOC ambiguity

Machine translation

NER

I love amazon, it's a great website
I love amazon, it's a beautiful river } BiRNN

Bidirectional RNN Architecture

26 October 2023 15:19

Amazon the best website.

Amazon the beautiful river.

loc / org

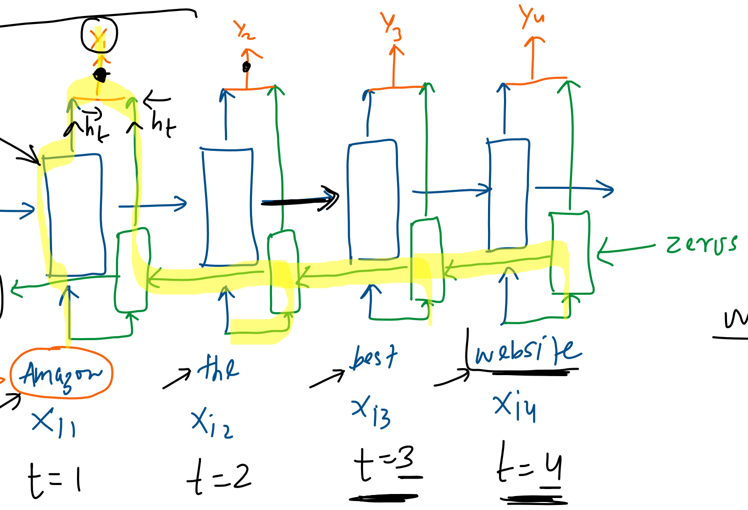
2 RNNs

✓ forward $\overrightarrow{RNN}$ - $\overleftarrow{RNN}$

concatenate

RNN

Zeros



$$\overrightarrow{h_t} = \tanh(\overrightarrow{W}\overrightarrow{h_{t-1}} + \overrightarrow{U}x_t + \overrightarrow{b})$$

$$\overleftarrow{h_t} = \tanh(\overleftarrow{W}\overleftarrow{h_{t+1}} + \overleftarrow{U}x_t + \overleftarrow{b})$$

$$\boxed{y_t = \sigma(V[\overrightarrow{h_t}, \overleftarrow{h_t}] + b)}$$

website → org

locx

Code

26 October 2023 15:21

{ BiLSTM } BiRNN

Concate